Universität Augsburg
Institut für Informatik

UX-orientierte Entwicklung von Empfehlungssystemen für Beratung und Assistenz

Dissertation

zur Erlangung des akademischen Grades des

Doctor rerum naturalium

(Dr. rer. nat.)

eingereicht an der Fakultät für Angewandte Informatik
Lehrstuhl Human Centered Multimedia

von

M.Sc. Stephan Martin Hammer

Augsburg, Deutschland, 2018

Gutachter:	Prof. Dr. Elisabeth André Prof. Dr. Jörg Hähner
Tag der mündlichen Prüfung:	24. Oktober 2018

Zusammenfassung

Begründet durch das Wachstum des Internets und der damit verbundenen Informationsflut haben Empfehlungssysteme in den letzten 25 Jahren immer weiter an Bedeutung gewonnen. In kommerziellen Anwendungsszenarios helfen sie Nutzern unter der Vielzahl an Objekten (z.B. Produkte, Filme, Immobilien, interessante Orte), die für sie nützlichsten zu finden. Über die Jahre wurden die Techniken zur Empfehlungsauswahl immer ausgereifter und auch die User Experience (UX) mit Qualitätskriterien wie Nutzerzufriedenheit, Nutzervertrauen und Nutzerakzeptanz spielte eine immer wichtigere Rolle in der Forschung.

Empfehlungssysteme könnten Menschen allerdings auch in alltäglichen Fragen beratend und assistierend zur Seite stehen. Durch proaktive Empfehlungen für Handlungen und Maßnahmen könnten Personen u.a. beim Umgang mit Krankheiten, bei der Steigerung ihres Wohlbefindens oder beim Schutz der Umwelt unterstützt werden. Hierfür müssen Empfehlungssysteme allerdings mit einer Reihe von Funktionalitäten und Fähigkeiten versehen werden, durch die eine möglichst gute Empfehlungsauswahl, aber auch eine gute UX (z.B. Überzeugungskraft, Nutzervertrauen, Nutzerakzeptanz) erreicht werden kann. In dieser Dissertation werden Ansätze vorgestellt, die neben der Empfehlungsauswahl auch die Generierung natürlichsprachlicher Empfehlungstexte (Argumentation und Formulierung) und die proaktive Ausführung von Empfehlungen betreffen.

Eine große Herausforderung für assistierende Empfehlungssysteme ist, dass häufig zu wenig Wissen über das Verhalten und die Werte der individuellen Personen zur Verfügung steht, um fundierte situative Entscheidungen über das Verhalten des jeweiligen Systems treffen zu können. Deswegen werden in dieser Dissertation geeignete sozialwissenschaftliche Modelle und Theorien, die dieses fehlende Wissen bereitstellen können, in Verfahren zur Entscheidungsfindung integriert. Für die Empfehlungsauswahl nutzen die in dieser Arbeit untersuchten Filterverfahren neben den klassischen Bewertungsmodellen auch theoriebasierte Nutzermodelle, die das Wohlbefinden und das Energieverhalten von Personen modellieren. Um überzeugende, personalisierte und situativ angepasste Empfehlungstexte generieren zu können, werden Kulturmodelle, Höflichkeitstheorien und die Persönlichkeit der Nutzer berücksichtigt. Für die Entscheidung, ob ein beratendes Empfehlungssystem zur Entlastung der Nutzer auch autonom gewisse Handlungen übernehmen darf, wird das Nutzervertrauen modelliert und in den Entscheidungsprozess miteinbezogen.

Zu den wissenschaftlichen Beiträgen dieser Dissertation zählen sowohl konzeptuelle Lösungsansätze als auch praktische Verfahren zur Entwicklung beratender Empfehlungssysteme. Ergänzt durch die während dieser Arbeit gewonnenen Erfahrungen, die ebenfalls beschrieben werden, soll Entwicklern und Forschern, die sich mit beratenden Empfehlungssystemen beschäftigen wollen, eine Grundlage geboten werden, um Systeme mit einer guten UX entwerfen und entwickeln zu können.

Abstract

Due to the growth of the Internet and the associated information flood, recommender systems have become increasingly important over the last 25 years. In commercial applications, recommender systems support users in finding the most useful among a variety of objects (e.g. products, movies, real estate, places of interest). Over the years, filter techniques have been constantly improved and user experience (UX) criteria such as user satisfaction, trust and acceptance have become increasingly important in research.

Recommender systems, however, can also assist people in everyday matters in an advisory and assisting manner. For example, proactive recommendations for actions could help people in dealing with diseases, increasing well-being or protecting the environment. To this end and in order to produce the best possible recommendation selection as well as a good UX (e.g., persuasiveness, user trust, user acceptance) recommender systems need additional functionalities and capabilities. In this thesis, approaches are presented, which in addition to the filtering of recommendations also consider the generation of natural language recommendation texts (e.g., reasoning and formulation) and the proactive execution of actions.

A major challenge for assistive recommender systems is the lack of knowledge about individual users' behavior and values, which is needed to enable informed system decisions and achieve improved situational system behavior. Therefore, this thesis integrates appropriate social-scientific models and theories that can provide the required knowledge into decision making processes. For the selection of recommendations, the approaches examined in this dissertation do not use only traditional rating based models but also theory based user models that model the well-being and energy behavior of individual users. In order to be able to generate persuasive, personalized and situational recommendation texts, cultural models, politeness theories and the personality of the users are taken into account. User trust is modeled and included in the decision-making process in order to decide whether an assistive recommender system is also allowed to perform certain actions autonomously.

The contributions of this doctoral thesis include conceptual as well as practical solutions for the development of advisory recommender systems. Supplemented by the experiences gained during this thesis work, which are also described in detail, the contributions of this thesis form a basis for developers and researchers who want to design and develop successful assistive recommender systems.

Danksagungen

Zunächst möchte ich mich herzlich bei meiner Doktormutter Prof. Dr. Elisabeth André bedanken. An ihrem ausgezeichneten Lehrstuhl hatte ich viele lehrreiche, spannende und schöne Jahre. Vor allem aber danke ich ihr für die vielen hilfreichen Kommentare und Ratschläge, die dazu beigetragen haben, dass sich meine Arbeit, aber auch ich mich selbst über den ganzen Zeitraum stetig weiterentwickelt haben.

Mein weiterer Dank gebührt Prof. Dr. Jörg Hähner, mit dem ich zunächst im Rahmen des OC Trust-Projekts zusammenarbeiten durfte und der sich anschließend auch dazu bereiterklärt hat, mein Zweitgutachter zu werden. Bei Prof. Dr. Dr.-Ing. Wolfgang Minker bedanke ich mich ebenfalls für seine Bereitschaft meine Dissertation zu begutachten.

Natürlich bedanke ich mich auch bei all meinen Kolleginnen und Kollegen am HCM-Lehrstuhl, allen Kollegen an anderen Lehrstühlen und Instituten sowie bei den unzähligen Studenten, mit denen ich zusammenarbeiten durfte. Ein besonderer Dank geht an Karin Bee, die mir als Betreuerin meiner Abschlussarbeiten, aber auch als Kollegin und Freundin den Start in mein Forscherleben erleichtert hat. Danke auch an meine (Zimmer-)Kollegen Katia Kurdyukova, Ilhan Aslan und Gregor Mehlmann für die vielen schönen und lustigen Stunden, die kreativen Gespräche und die gute Zusammenarbeit in den Projekten und Vorlesungen. Auch bei Birgit Lugrin, Michael Wissner sowie Andreas Seiderer bedanke ich mich für die tolle, kreative Zusammenarbeit im Rahmen der verschiedenen Projekte.

Abseits der Universität möchte ich mich zuerst bei meinen Freunden bedanken, die mich durch ihre Teilnahme an den verschiedenen Studien, ihre Hilfe beim Korrektur lesen, aber auch allein durch ihre Freundschaft unterstützt haben. Auch meiner Schwiegermama möchte ich für ihre Hilfe beim Korrektur lesen danken.

Sehr dankbar bin ich meinen Großeltern und Eltern. Sie haben mich zu dem Menschen gemacht, der ich heute bin, und mir diesen Weg damit erst ermöglicht. Speziell meinen Eltern möchte ich von ganzem Herzen dafür danken, dass sie mir über all die Jahre hinweg den nötigen Rückhalt und Antrieb gegeben haben.

Am allermeisten möchte ich mich bei meiner wunderbaren Frau Katrin bedanken. Vielen Dank, dass du mich in all den Jahren unterstützt hast. Vielen Dank, dass du dir alles Positive, aber auch alles Negative angehört und mir deine Meinung dazu gesagt hast. Vielen Dank für deine Geduld in all der Zeit und, dass du mir im letzten Jahr seit Sophia unser Leben bereichert, immer wieder den Rücken freigehalten hast, damit ich meine Arbeit abschließen konnte.

Auch meiner kleinen Maus möchte ich danken. Du hast mir im Endspurt nochmal viel Kraft und Motivation gegeben.

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Empfehlungssysteme und ihre Anwendungsszenarios Anfang bis Mitte der 90er-Jahre entstanden die ersten *Empfehlungssysteme* (engl. Recommender Systems), siehe z.B. [Goldberg et al., 1992, Resnick et al., 1994]. Aufgrund der wachsenden Daten- und Informationsflut im schnell wachsenden Internet war es immer wichtiger geworden, die Nutzer bei der Suche nach für sie interessanten Informationen und Objekten zu unterstützen.

Diese Aufgabe erledigen Empfehlungssysteme, indem sie ihren Nutzern basierend auf einem Nutzermodell in einer gegebenen Situation, die durch den Kontext beschrieben wird, aus einer Gesamtmenge an Objekten mit Hilfe von Filteralgorithmen eine Teilmenge an Objekten empfehlen, die für die Zielperson(en) den größten Nutzen haben [Resnick und Varian, 1997]. Die *Nutzermodelle* enthalten in den meisten Fällen die von den Nutzern abgegebenen Bewertungen für bereits bekannte Objekte. Je nach Filteralgorithmus können allerdings u.a. auch demographische Daten oder aber zuletzt gekaufte oder genutzte Objekte berücksichtigt werden. Schematisch ist das Zusammenspiel der einzelnen Komponenten in Abbildung 1.1 dargestellt.

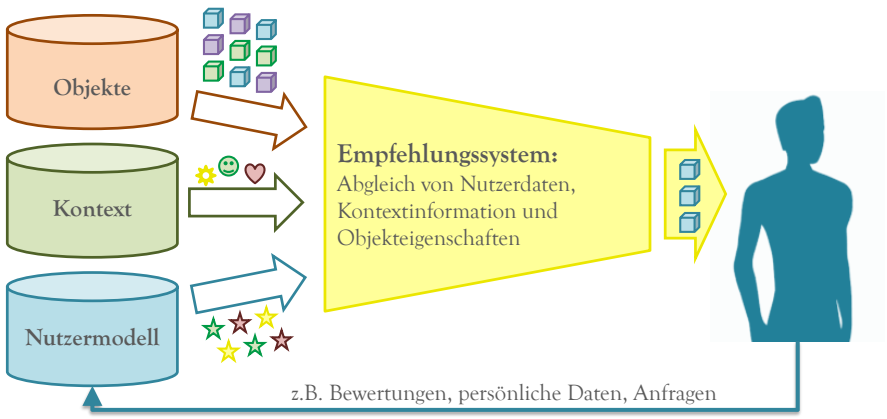

Abbildung 1.1: Empfehlungssysteme - Genereller Aufbau und Funktionsweise

Empfehlungssysteme wurden sowohl in der Forschung als auch im kommerziellen Bereich schnell populär. Ein bekanntes Beispiel für die kommerzielle Nutzung eines Empfehlungssystems ist Amazon[1]. Aber auch zahlreiche weitere Unternehmen machen sich die Dienste von Empfehlungssystemen zu Nutze. Beispiele sind Netflix[2] zur Empfehlung von Filmen und Serien, ImmobilienScout24[3] zur Suche nach Immobilien oder TripAdvisor[4] zur Suche nach Restaurants oder Sehenswürdigkeiten.

Auch in anderen Anwendungsbereichen werden Empfehlungssysteme immer häufiger erforscht und eingesetzt. Zum Beispiel wurden Systeme entwickelt, die Menschen bei der Suche nach Ärzten [Hoens et al., 2013] oder Schulen für ihre Kinder

[1] http://www.amazon.de
[2] http://www.netflix.com/de
[3] https://www.immobilienscout24.de/
[4] https://www.tripadvisor.de/

[Wilson et al., 2009] unterstützen. Des Weiteren gibt es auch Systeme zur Förderung einer gesunden Ernährung [van Pinxteren et al., 2011] und körperlicher Aktivitäten [Lin et al., 2011]. Felfernig und Kollegen [Felfernig et al., 2013] bezeichnen diese Art von Empfehlungssystemen als *persönliche Assistenten*.

Beratende und assistierende Empfehlungssysteme zu entwickeln, ist auch das Ziel dieser Dissertation. Die untersuchten Systeme sollen in das Leben ihrer Nutzer integriert werden und ihnen über den Tag verteilt proaktiv Empfehlungen für Aktivitäten und Maßnahmen aussprechen, um sie bei der Bewältigung alltäglicher Herausforderungen zu unterstützen. Ein erstes, richtungsweisendes Projekt zu Beginn dieser Dissertation zielte darauf ab, Diabetikern durch Empfehlungen für Sportaktivitäten und eine angepasste Ernährung den Umgang mit ihrer Krankheit zu erleichtern und dadurch möglicherweise sogar die Menge der benötigten Medikation zu reduzieren [Hammer et al., 2010]. Die Anwendungsszenarien, die für die Untersuchungen in dieser Dissertation verwendet wurden, befassen sich mit der Unterstützung alleinstehender Senioren sowie mit der Beratung beim Energiesparen. Sie werden in Kapitel 1.2 ausführlich vorgestellt.

Eine große Herausforderung bei der Entwicklung assistierender Empfehlungssysteme ist, dass die Nutzer die „Einmischungen“ der Systeme in ihre Leben akzeptieren und die Empfehlungen annehmen müssen. Ob dies gelingt, hängt stark von den individuellen Nutzern und ihrer Wahrnehmung des Systems während der Nutzung ab. Aus diesem Grund liegt der Schwerpunkt dieser Dissertation auf der *User Experience (UX)* beratender Empfehlungssysteme.

Empfehlungssysteme und ihre User Experience Nach der allgemeinen Definition [Deutsches Institut für Normung e.V., 2010] umfasst die *User Experience (UX)* einer Person ihre Wahrnehmung und Reaktion, die sich bei der Benutzung oder der erwarteten Verwendung eines Produkts, Systems oder Dienstes ergibt. Wichtige Faktoren der UX sind zum Beispiel die Ansichten der Person sowie ihre psychologischen und physiologischen Reaktionen, Verhaltensweisen und Leistungen vor, während und nach der Nutzung.

Im Bereich Empfehlungssysteme ging man lange davon aus, dass hauptsächlich durch die Entwicklung und Verbesserung von Filteralgorithmen eine gute UX erreicht werden könnte. Genauere Empfehlungen sollten für die Nutzer nützlicher sein und ihnen auch besser gefallen. Mit der Zeit kam man jedoch zu der Erkenntnis, dass auch andere Kriterien wie die Vielfalt der Empfehlungen, die Möglichkeit der Nutzer, auf die Empfehlungsergebnisse einzuwirken sowie die wahrgenommene Transparenz der Systeme maßgeblich die UX von Empfehlungssystemen beeinflussen [Jannach et al., 2016, Knijnenburg et al., 2012, Konstan und Riedl, 2012, McNee et al., 2006, McNee, 2006].

Konstan und Riedl [Konstan und Riedl, 2012] unterteilten die UX mit Empfehlungssystemen in die Präsentation von Empfehlungen und die Interaktion der Nutzer mit den Empfehlungen. Knijnenburg und Kollegen [Knijnenburg et al., 2012] unterschieden ähnlich zur allgemeinen Definition von UX

[Deutsches Institut für Normung e.V., 2010] die UX vor, während und nach dem Erhalt einer Empfehlung. Wichtige Faktoren in diesen Phasen sind u.a. der wahrgenommene Aufwand für die Nutzer bis zum Erhalt einer Empfehlung, die wahrgenommene Effektivität des Systems während der Präsentation der Empfehlungen und die Nutzerzufriedenheit als Ergebnis des kompletten Empfehlungsprozesses.

Für die in dieser Arbeit untersuchten Anwendungsszenarien ergeben sich zusätzlich zu den genannten Faktoren weitere Herausforderungen für die UX. Betrachtet man das angesprochene Empfehlungssystem für Diabetiker zeigt sich schon alleine durch das Aussprechen der Empfehlungen ein gewisses Konfliktpotential. Werden weniger aktiven oder sich ungesund ernährenden Personen Sportübungen oder gesunde Rezepte empfohlen, können diese sich ertappt und peinlich berührt oder sogar bevormundet fühlen. Diese und weitere Gefahren für die Beziehung zwischen System und Nutzer lassen sich auch in vielen anderen Anwendungsszenarien finden, siehe auch Kapitel 1.2. Demzufolge stellt sich in dieser Dissertation die Frage, wie beratende Empfehlungssysteme agieren müssen, damit die genannten Risiken minimiert werden und die Nutzer gerne mit ihnen interagieren.

Mit der Frage, wie Computertechnologien gestaltet sein sollten, damit sie bei Nutzern Akzeptanz finden, hat sich Friedman [Friedman, 1996, Friedman und Kahn, 2003] beschäftigt. Der Kern ihrer Forschung befasste sich damit, welche menschlichen Werte sich wie im Design von Systemen widerspiegeln sollten. Werte, die laut Friedman berücksichtigt werden sollten, sind u.a. die Sicherung der eigenen Privatsphäre, die Autonomie der Nutzer und das Vertrauen in die Fähigkeiten und das Wohlwollen eines Interaktionspartners. Diese Werte haben laut Pu und Kollegen [Pu et al., 2011] auch einen starken Einfluss auf die UX von Empfehlungssystemen. Sie ergänzen u.a. noch die Transparenz der Systemaktionen, die Nützlichkeit der Systeme und die Einfachheit der Nutzung. Im Zusammenhang mit allgegenwärtigen Systemen ist außerdem eine gewisse Zurückhaltung der Systeme von Bedeutung, damit sie sich auf angemessene Weise in den Alltag der Nutzer integrieren [Friedman, 1996, Friedman und Kahn, 2003]. In dieser Dissertation wird erforscht, wie die genannten Werte bei der *Empfehlungsauswahl*, der *Generierung von Empfehlungstexten* sowie bei der *autonomen Ausführung von Empfehlungen* berücksichtigt werden können, um für beratende Empfehlungssysteme eine gute UX erreichen zu können.

Um den Erfolg der in dieser Arbeit entwickelten Verfahren messen zu können, müssen Qualitätskriterien festgelegt werden, anhand derer die UX assistierender Empfehlungssysteme bewertet werden kann.

In ihrem Framework zur nutzerzentrierten Evaluation von Empfehlungssystemen nannten Pu und Kollegen [Pu et al., 2011] die Akzeptanz eines Systems bzw. die Intention, das System zu nutzen, sowie die Intention, den Empfehlungen des Systems zu folgen, als die Hauptkriterien, anhand derer Empfehlungssysteme bewertet werden sollten. Diese Kriterien werden in dieser Dissertation unter den Begriffen *Überzeugungskraft* und *Nutzerakzeptanz* zusammengefasst. Als zusätzliches Qualitätskriterium wird außerdem die *Vertrauenswürdigkeit* bzw. das Vertrauen der

Nutzer in ein beratendes Empfehlungssystem hinzugezogen. Dieses Kriterium ist deswegen von großer Bedeutung, da die untersuchten Systeme Maßnahmen und Aktivitäten empfehlen, die von den Nutzern einen gewissen Aufwand und möglicherweise auch Überwindung erfordern. Außerdem sollen die Systeme zum Teil auch autonom Maßnahmen ausführen können. In beiden Fällen müssen sich die Nutzer auf die Korrektheit und das Wohlwollen der Systeme verlassen.

1.2 Anwendungsszenarien

Beratende und assistierende Empfehlungssysteme sind in verschiedensten Anwendungsszenarien denkbar. In dieser Dissertation wurden als Beispiel Prototypen für zwei wichtige, gesellschaftliche Themen untersucht: (1) Förderung des Wohlbefindens alleinstehender Senioren und (2) Förderung energiesparenden Verhaltens. Diese beiden Anwendungsszenarien wurden ausgewählt, weil sie unterschiedliche Herausforderungen für die Gestaltung der jeweiligen Systeme mit sich bringen.

1.2.1 CARE (Context-Aware Recommender system for the Elderly)

Der Fokus des CARE-Szenarios liegt auf alleinstehenden Senioren. Diese verlieren häufig aufgrund körperlicher und mentaler Einschränkungen an Selbstständigkeit und sind deswegen in manchen Bereichen des Alltags auf Hilfe von Außen angewiesen. Weiterhin führen die Einschränkungen oft auch zu Inaktivität und fehlender Selbstinitiative. Im schlimmsten Fall kommt es sogar zur sozialen Isolation.

Durch gezielt ausgewählte Aktivitäten ist es allerdings möglich, diesen Gefahren entgegenzuwirken. Regelmäßige körperliche Aktivitäten können erwiesenermaßen die Folgen vieler altersbedingter Krankheiten wie Rheuma oder Diabetes mildern [Teri und Lewinsohn, 1982, Mahncke et al., 2006]. Auch kreative und geistige Aktivitäten wie Rätsel und Gedächtnisübungen, aber auch Malen oder Gartenarbeit, können das Wohlbefinden steigern [Schmid, 2005]. Des Weiteren konnte in Studien gezeigt werden, dass soziale Aktivitäten wie das Übernehmen von Aufgaben in einer häuslichen Gemeinschaft einen beträchtlichen Einfluss bei der Vermeidung sozialer Isolation und Einsamkeit haben [Arnetz und Theorell, 1983].

Es gibt bereits eine Vielzahl an Arbeiten, die einen gesunden Lebensstil, die Lebensqualität oder das Wohlbefinden der Nutzer fördern wollen. Einige der Systeme arbeiten mit personalisiertem Feedback in Form von Metaphern wie Aquarien oder Gärten, die die Vitalität ihrer Nutzer widerspiegeln sollen [Consolvo et al., 2008, Lane et al., 2014]. Andere Systeme versuchen auf einfache Weise, einen direkten kausalen Zusammenhang zwischen dem aktuellen Lebensstil (z.B. Schlafgewohnheiten, körperliche Aktivitäten) und dem persönlichen Wohlbefinden herzustellen [Bentley et al., 2013]. Das Motivate-System nutzte Kontextinformation (z.B. Position der Nutzer, Wetter) und den aktuellen Zeitplan der Nutzer, um zum richtigen Zeitpunkt und am richtigen Ort situative Empfehlungen für körperliche Aktivitäten auszusprechen [Lin et al., 2011]. Auch virtuelle Agenten [Ring et al., 2015] und

Roboter [Turunen et al., 2008] wurden bereits als Ratgeber für einen gesunden Lebensstil eingesetzt. Nur wenige dieser Systeme (z.B. [Ring et al., 2015]) fokussierten allerdings auf das Wohlbefinden und die Bedürfnisse älterer Menschen.

Das CARE-System hat das Ziel durch proaktive Empfehlungen den Anstoß zur Durchführung von Aktivitäten zu geben, die das Wohlbefinden älterer Menschen fördern können. Dadurch sollen diese wieder motivierter, belastbarer und selbstbewusster werden. Auf lange Sicht könnte den Senioren so ermöglicht werden, länger relativ selbstständig und vor allem in ihrem eigenen Zuhause zu leben.

Eine große Herausforderung für das CARE-System ist der Gewinn der Nutzerakzeptanz. Bis heute steht ein großer Teil der Senioren neuen Technologien skeptisch bis ablehnend gegenüber. Das System muss diese Nutzer also davon überzeugen, dass es einen tatsächlichen Nutzen für sie bieten kann. Ein weiteres Problem, das speziell im CARE-Szenario auftreten kann, ist die Entstehung emotionaler Barrieren. Das System wird die Senioren durch seine Empfehlungen nicht nur auf vorhandene Schwächen oder Fehlverhalten hinweisen. Es wird ihnen auch noch „kluge Ratschläge“ geben, was sie besser machen könnten. Das wiederum ist ein Problem, das selbst zwischen Menschen häufig zu Konflikten führt.

Die Forschung hinsichtlich des Anwendungsszenarios CARE wurde im Rahmen des namensgebenden BMBF-Projektes CARE[5] und innerhalb des vom Bayerischen Staatsministerium für Bildung und Kultus, Wissenschaft und Kunst geförderten Forschungsverbundes ForGenderCare[6] durchgeführt.

1.2.2 SavER (SituAtiVes Energie Recommender system)

Den individuellen Energieverbrauch zu senken, ist seit längerem Teil vieler Forschungsarbeiten [DiSalvo et al., 2010, Hazas et al., 2011]. Manche Arbeiten beschäftigten sich mit der Anzeige detaillierten Feedbacks bzgl. des persönlichen Energieverbrauchs der Nutzer [Froehlich et al., 2009, Gamberini et al., 2012]. Andere konzentrierten sich auf soziale Faktoren und erforschten zum Beispiel kooperative und überzeugende Spiele, in denen energiesparendes Verhalten belohnt wird [Bühling et al., 2012, Simon et al., 2012]. Auch Hausautomatisierungssysteme sind ein wichtiger Ansatz. Sie erlauben es u.a. , Steckdosen, Lichter oder Heizungen entweder aus der Ferne oder basierend auf Zeitplänen zu steuern. Intelligente Varianten dieser Systeme nutzen Kontextinformationen, um selbstständig Geräte steuern zu können [Cheverst et al., 2005].

Ein Hinderungsgrund für Verhaltensänderungen ist häufig, dass vielen Menschen das nötige Wissen fehlt, selbst tätig zu werden. Sie benötigen konkrete und einfach zugängliche Tipps, mit welchen Maßnahmen sie Energie sparen können [Gardner und Stern, 2008]. Auf Infoseiten wie denen des WWF [WWF Deutschland, 2016] oder des Bundesministerium für Umwelt [Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit, 2016] können sich Interessierte informieren. Außerdem gibt es Webseiten,

[5] http://care-project.net

[6] http://www.forgendercare.de

auf denen man seinen eigenen Verbrauch analysieren lassen kann [RheinEnergie AG, 2017, co2 online, 2017]. Jedoch liefern die meisten dieser Ratgeber nur sehr allgemeine Tipps, die nicht oder kaum auf die Nutzer, ihre Möglichkeiten und ihr bisheriges Verhalten zugeschnitten sind. Dadurch sind Systeme dieser Art nur selten wirksam.

Um sicherzustellen, dass Energiespartipps angewendet werden, müssten sie personalisiert werden [Benders et al., 2006]. Shigeyoshi und Kollegen [Shigeyoshi et al., 2013] entwickelten ein entsprechendes System. Allerdings orientierte sich die Empfehlungsauswahl ihres Systems zu stark an der Effektivität der Energiesparaktionen. Die Präferenzen der Nutzer für Aktionen sowie ihre Möglichkeiten zur Durchführung der Aktionen wurden nicht berücksichtigt, so dass weiterhin nur wenige Empfehlungen angenommen wurden. Weitere Untersuchungen wurden von Ford und Kollegen [Ford et al., 2014] durchgeführt. Bei ihnen basierte die Empfehlungsauswahl auf den Präferenzen der Nutzer für bestimmte Eigenschaften der Maßnahmen. Hierzu gehörten u.a. der finanzielle, zeitliche oder körperliche Aufwand sowie die Zuverlässigkeit der Maßnahmen. Die Nutzer des Webportals erhielten aber nur Empfehlungen, wenn sie aktiv danach suchten. Oft fehlt vor allem Nutzern, die bisher nur wenig aktiv waren, hierzu jedoch die Motivation. Zusätzlich geht die Chance auf weitere, spontane Energieersparnisse verloren. Zu guter Letzt wurden im Portal von Ford und Kollegen weder das tatsächliche Verhalten, noch frühere Entscheidungen der Nutzer bei der Empfehlungsauswahl berücksichtigt.

Das SavER-System soll die Schwächen bisheriger Systeme beseitigen und den Nutzern in ihrem Alltag proaktiv energiesparende Maßnahmen empfehlen. Dabei sollen auch das aktuelle Energieverhalten der Nutzer sowie die aktuelle Situation berücksichtigt werden.

Als besondere Herausforderung im SavER-Szenario lassen sich die unterschiedliche Motivation der Nutzer für energiesparendes Verhalten sowie eine Überreizung durch zu viele Empfehlungen hervorheben. Die unterschiedlichen Beweggründe erschweren die Auswahl überzeugender Argumente, die mit den Empfehlungen angezeigt werden sollten, um die Nutzer auch zu aufwendigeren und ungewohnten Handlungen zu motivieren. Eine Überreizung durch zu viele Empfehlungen könnte vor allem dann auftreten, wenn vermehrt kleinere Maßnahmen wie das Ausschalten von Lichtern oder Geräten vorgeschlagen werden. Anstatt die Nutzer durch zu häufige Empfehlungen zu belästigen, könnte das SavER-System solche einfachen Maßnahmen in bestimmten Situationen selbst ausführen und seine Nutzer somit entlasten.

Die Untersuchungen im Anwendungsszenario SavER fanden im Rahmen der Projekte OC-Trust (DFG-gefördert)[7] und IT4SE (BMBF-gefördert)[8] statt.

[7] `http://www.isse.uni-augsburg.de/en/projects/reif/oc-trust/`

[8] `http://it4se.hs-augsburg.de/`

1.3 Herausforderungen & Forschungsfragen

Ein entscheidender Bestandteil assistierender Empfehlungssysteme sind *Nutzermodelle*. Nutzermodelle werden in der Mensch-Computer-Interaktion (engl. Human-Computer-Interaction HCI) und der künstlichen Intelligenz eingesetzt, damit sich Systeme in irgendeiner Form an die Nutzer und ihre individuellen Bedürfnisse anpassen und somit ihre Funktionalität personalisieren können [Jameson, 2003].

In diesem Kapitel werden für die in dieser Dissertation behandelten Themengebiete (Empfehlungsauswahl, Generierung von Empfehlungstexten und autonome Ausführung von Empfehlungen) die Herausforderungen und Forschungsfragen für die Nutzermodellierung bzw. für den Einsatz von Nutzermodellen beschrieben.

1.3.1 Empfehlungsauswahl

Um die Chance für die Durchführung empfohlener Aktivitäten und Maßnahmen zu erhöhen, müssen diese für die Zielperson im Moment der Empfehlung von Interesse und Nutzen sein. Deshalb müssen die Systeme einerseits zur Laufzeit die Nutzerbewertungen für die möglichen Empfehlungen korrekt vorhersagen und die tatsächliche Nützlichkeit der Aktionen und Maßnahmen in der vorliegenden Situation einschätzen können. Andererseits müssen auch bereits bei der Entwicklung der Systeme die Bedürfnisse und Anforderungen der Nutzer berücksichtigt werden.

Traditionelle Verfahren zur Empfehlungsauswahl wie das kollaborative Filtern, siehe Kapitel 3.2, haben häufig Probleme damit, relevante Objekte zu identifizieren, wenn Nutzer zuvor nur wenige Bewertungen abgegeben haben. In beratenden Empfehlungssystemen kommt erschwerend hinzu, dass manche Empfehlungen nur Sinn machen, wenn sie sich auf die aktuellen Fähigkeiten, Meinungen oder Verhaltensweisen der Nutzer beziehen. Hat eine ältere Person zum Beispiel akute Knieschmerzen, hat eine Empfehlung zum Spazierengehen aktuell keinen Nutzen für sie. Abgegebene Bewertungen für Empfehlungen können sich demzufolge von Situationen zu Situation stark unterscheiden [McNee, 2006]. Um dennoch gute Empfehlungen auswählen zu können, benötigen die Systeme also zusätzliches Wissen.

In dieser Arbeit wurde die Idee verfolgt, für die Empfehlungsauswahl fundierte sozialwissenschaftliche Theorien und Modelle hinzuzuziehen, mit denen die aktuellen Fähigkeiten, Werte und Verhaltensweisen der Nutzer modelliert werden können. Im CARE-System würde zum Beispiel ein Modell benötigt, das u.a. das aktuelle körperliche und mentale Wohlbefinden der Senioren beschreibt. Im SavER-System wären dagegen Informationen über die aktuelle Einstellung zum Thema Energiesparen, das tatsächliche Verhalten und Möglichkeiten zum Energiesparen hilfreich.

Forschungsfragen Wie lassen sich theoriebasierte Nutzermodelle in das klassische kollaborative Filterverfahren integrieren? Kann durch die Integration der theoriebasierten Nutzermodelle die Qualität der Empfehlungen in Situationen mit wenigen Bewertungen verbessert werden?

Die Wahrnehmung der Empfehlungsauswahl eines Systems hängt jedoch nicht nur von der Korrektheit der Empfehlungen ab. Faktoren wie die Anzahl der Empfehlungen oder die Möglichkeit zur Beeinflussung der Empfehlungsauswahl wirken sich ebenfalls auf die UX der Systeme aus [Jannach et al., 2016, Knijnenburg et al., 2012, Konstan und Riedl, 2012, McNee et al., 2006, McNee, 2006].

Am Beispiel des CARE-Szenarios wird in dieser Dissertation die nutzerzentrierte Entwicklung eines beratenden Empfehlungssystems vorgestellt. Da Senioren häufig skeptisch gegenüber neuen Technologien sind, sollte bei der Entwicklung besonders auf ihre Wünsche und möglicherweise auch Ängste eingegangen werden, damit das System später akzeptiert wird. Die einzelnen, in dieser Arbeit beschriebenen Schritte des Entwicklungsprozesses von einer Anforderungsanalyse über die Evaluation eines ersten Prototypen bis zur Umsetzung eines überarbeiteten zweiten Prototypen zeigen, welche besonderen Anforderungen und Bedürfnisse die befragten Nutzer an das CARE-System hatten.

1.3.2 Generierung von Empfehlungstexten

Neben einer qualitativ hochwertigen Empfehlungsauswahl spielen in assistierenden Empfehlungssystemen auch der Inhalt und die Formulierung der Empfehlungstexte eine wichtige Rolle für die UX.

Zunächst müssen die Nutzer verstehen, wieso sie eine bestimmte Empfehlung erhalten und vor allem warum sie diese annehmen sollten [Jannach et al., 2016, Sinha und Swearingen, 2002]. Deshalb wird in dieser Dissertation auch die Überzeugungskraft der Argumentation in Empfehlungstexten untersucht.

Die Beweggründe für die Befolgung einer Empfehlung und auch die Wahrnehmung bestimmter Argumente und Überzeugungsstrategien unterscheiden sich jedoch von Mensch zu Mensch. Ausschlaggebend sind u.a. die individuellen Werte, Ziele und Bedürfnisse. Es ist allerdings nur schwer möglich, über jede Person ausreichend detaillierte und aktuelle Daten zu sammeln, anhand derer personalisierte Argumente ausgewählt werden könnten.

Die Werte, Ziele und Bedürfnisse der Menschen werden jedoch auch stark durch ihre Kultur beeinflusst. Welche Werte und Eigenschaften bestimmte Kulturgruppen auszeichnen, ist in bekannten Kulturmodellen beschrieben [Hofstede, 2001, Hofstede et al., 2010, Triandis, 1995]. So zeichnen sich US-Amerikaner zum Beispiel durch ein stark individualisiertes Denken aus, während Chinesen stark kollektivistisch eingestellt sind [Hofstede, 2017]. Dieses Wissen könnte man in beratenden Empfehlungssystemen nutzen, um anhand der in einer Kultur typischen Werte und Eigenschaften die Argumente auszuwählen, die für die Menschen dieser Kulturgruppe als am überzeugendsten gelten. Für US-Amerikaner könnten das Argumente sein, die nur sie und ihr enges Umfeld betreffen. Chinesen könnten dagegen durch Argumente überzeugt werden, die sie und ihre Stadt oder ihr Land betreffen.

Forschungsfrage - Kulturbasierte Argumentauswahl Kann ein assistierendes Empfehlungssystem die Überzeugungskraft von Empfehlungstexten dadurch steigern, dass es die Argumente für eine empfohlene Maßnahme basierend auf typischen Werten und Eigenschaften der Kulturgruppe der Zielpersonen auswählt?

Neben dem argumentatorischen Inhalt spielt in beratenden Empfehlungssystemen auch die Formulierung der Empfehlungstexte eine wichtige Rolle. Die Formulierung „Geh mal wieder spazieren." hat eine andere Wirkung als die Formulierung „Wie wäre es, wenn du mal wieder spazieren gehst?". Ein Empfehlungstext sollte, falls möglich, so formuliert werden, dass er zwar überzeugend, aber weder bevormundend, noch bestimmend wirkt. Die Höflichkeit sowie die wahrgenommene Persönlichkeit eines Systems, aber auch die Persönlichkeit der Nutzer selbst haben zum Beispiel einen deutlichen Einfluss darauf, wie bestimmte Aussagen aufgenommen werden [Brown und Levinson, 1987, Nass und Lee, 2001, Reeves und Nass, 1998].

Forschungsfrage - Formulierung Kann durch die gezielte Formulierung eines Empfehlungstextes die individuelle Wahrnehmung einer Empfehlung beeinflusst werden? Welche Strategien eignen sich, um situativ eine gute Balance zwischen Höflichkeit und Überzeugungskraft zu erreichen? Wirkt eine Empfehlung vertrauenswürdiger und überzeugender, wenn sie die Persönlichkeit der Zielperson widerspiegelt?

1.3.3 Proaktives Ausführen von Empfehlungen

Sprechen Empfehlungssysteme zu häufig und womöglich sogar in unpassenden Situationen Empfehlungen aus, können diese als störend und lästig wahrgenommen werden [Bader et al., 2010, Melguizo et al., 2007]. Die Nutzer könnten außerdem das Gefühl haben, dass sie die Kontrolle über das System verlieren.

Abhängig vom jeweiligen System und der aktuellen Situation könnte es jedoch passieren, dass Maßnahmen wie das Ein- und Ausschalten von Geräten oder das Festlegen von Terminen auch dann durchgeführt werden sollten, wenn die Zielperson nicht (mehr) gestört werden sollte. Eine Lösung für dieses Problem wäre, dass das System die entsprechende Maßnahme, falls möglich, selbstständig durchführt. Allerdings muss dann geklärt werden, in welchen Situationen solch ein autonomes Handeln angebracht ist und in welchen nicht.

Ein Entscheidungskriterium könnte das Vertrauensverhältnis zwischen Nutzer(in) und System sein. Das Nutzervertrauen müsste allerdings so modelliert werden können, dass das jeweilige Empfehlungssystem situativ die Auswirkungen spezifischer Systemaktionen, wie dem Aussprechen einer Empfehlung, dem autonomen Handeln oder dem nicht aktiv werden, auf das Nutzervertrauen einschätzen kann.

Forschungsfrage Wie kann das Nutzervertrauen gegenüber einem beratenden Empfehlungssystem modelliert werden, so dass dem System situative Entscheidungen über die Angemessenheit verschiedener Systemaktionen ermöglicht werden?

Wie stark beeinflussen Faktoren wie Transparenz, Nutzerkontrolle oder Nutzungskomfort das Vertrauen der Nutzer sowie ihre Präferenzen gegenüber spezifischen Systemaktionen?

1.4 Ziel der Dissertation

Bei der Empfehlungsauswahl, der Generierung von Empfehlungstexten und bei der Einschätzung der Angemessenheit der proaktiven Ausführung von Empfehlungen sind beratende Empfehlungssysteme auf das Wissen aus fundierten Theorien über das Verhalten und die Werte der Menschen angewiesen. Es ist anzunehmen, dass nur mit Hilfe dieses Wissens die User Experience und dabei im Speziellen die Überzeugungskraft, die Nutzerakzeptanz und das Nutzervertrauen gesteigert werden können.

Die wissenschaftlichen Beiträge dieser Dissertation sind konzeptuelle und praktische Lösungsansätze, die die Integration sozialwissenschaftlicher Theorien und Modelle in die genannten Schritte des Empfehlungsprozesses ermöglichen. Entwickelt und untersucht werden diese Ansätze durch prototypische Umsetzungen in den Anwendungsszenarien CARE und SavER. Außerdem werden auch die während der Entwicklung und Evaluation der prototypischen Systeme gemachten Erfahrungen beschrieben, um anderen Forschern dabei zu helfen, die vorgeschlagenen Ansätze im Rahmen der Entwicklung ihrer eigenen assistierenden Empfehlungssysteme einzusetzen und gegebenenfalls an ihre Anforderungen anzupassen.

1.5 Struktur der Arbeit

Den beschriebenen Herausforderungen und Forschungsfragen wurden in dieser Dissertation eigene Kapitel gewidmet. Abbildung 1.2 zeigt den Zusammenhang zwischen diesen Kapiteln und den jeweiligen Faktoren der Nutzermodellierung. Die Reihenfolge der Forschungskapitel (Kapitel 4 bis 6) orientierte sich an der schrittweisen Erweiterung der Funktionalität beratender Empfehlungssysteme und nicht an der in Abbildung 1.2 dargestellten logischen Reihenfolge der Arbeitsschritte im Empfehlungsprozess. Die Kapitel sind jedoch für sich eigenständig aufgebaut, weswegen die Lesereihenfolge keinen Einfluss auf das Verständnis der Arbeit hat.

Zunächst werden in Kapitel 2 die Konzepte Vertrauen, Überzeugungskraft und Nutzerakzeptanz erklärt. Die Konzepte Wohlbefinden, Energieverhalten, Kultur, Persönlichkeit und Höflichkeit sind nur in einzelnen Kapiteln von Bedeutung und werden zum besseren Verständnis erst in den entsprechenden Kapiteln vorgestellt.

In Kapitel 3 werden theoretische Grundlagen im Bezug auf Empfehlungssysteme im Allgemeinen vermittelt. Dazu gehören zum einen traditionell häufig eingesetzte Filtertechniken wie das kollaborative Filtern und zum anderen verschiedene Evaluationsmetriken. Auch wenn einzelne der vorgestellten Techniken und Metriken in dieser Dissertation keine Verwendung findet, sind sie womöglich bei der Entwicklung anderer Systeme von Bedeutung und werden deswegen der Vollständigkeit halber ebenfalls kurz vorgestellt.

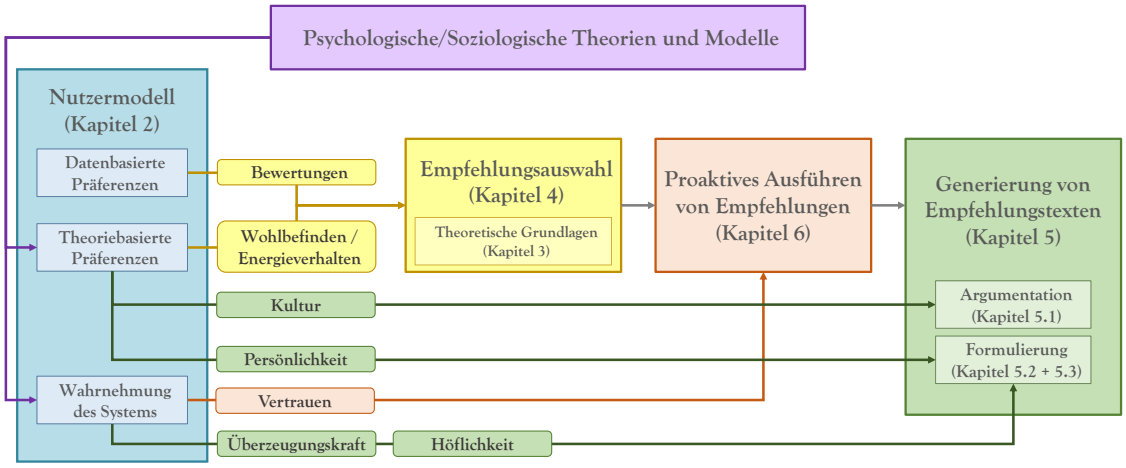

Abbildung 1.2: Struktureller Aufbau der Arbeit und logische Zusammenhänge und Abläufe innerhalb der untersuchten Empfehlungssysteme

Kapitel 4 geht auf die Empfehlungsauswahl ein und wird zunächst verwandte Arbeiten präsentieren, die erfolgreich theoriebasierte Nutzermodelle für die Filterung von Empfehlungen eingesetzt haben, siehe Kapitel 4.1. In Kapitel 4.2 werden theoriebasierte Nutzermodelle vorgestellt, die für die Empfehlungsauswahl in den Anwendungsszenarien CARE und SavER als geeignet erscheinen. Die Integration dieser Modelle in ein kollaboratives Empfehlungssystem sowie die Evaluation der neu entstandenen Filteransätze sind in Kapitel 4.3 beschrieben. Kapitel 4.4 beschreibt anschließend das Vorgehen und die Erkenntnisse innerhalb der nutzerzentrierten Entwicklung eines prototypischen CARE-Systems.

Kapitel 5 widmet sich der Generierung von Empfehlungstexten. Nachdem zunächst mittels verwandter Arbeiten ein Grundverständnis für Erklärungen und Argumente in Empfehlungssystemen vermittelt wird, wird in Kapitel 5.1 die Auswahl überzeugender, personalisierter Argumente untersucht. Der Fokus liegt auf kulturellen Unterschieden bei der Wahrnehmung von Argumenten. Da nicht nur der Inhalt eines Empfehlungstextes die Wahrnehmung der Nutzer beeinflusst, sondern auch die Formulierung des Textes, sind die wahrgenommene Höflichkeit von Empfehlungstexten, siehe Kapitel 5.2, sowie ihre wahrgenommene Persönlichkeit, siehe Kapitel 5.3, ebenfalls Bestandteil der Untersuchungen des fünften Kapitels.

Die Erweiterung der Funktionalität eines beratenden Empfehlungssystems durch die autonome Ausführung von Maßnahmen zur Entlastung der Nutzer wird in Kapitel 6 untersucht. In Kapitel 6.1 werden verwandte Arbeiten beschrieben, die Faktoren untersuchten, die Nutzervertrauen gegenüber adaptiven Systemen beeinflussen. Kapitel 6.2 beschreibt dann mit dem User Trust Model (UTM) einen Ansatz zur Modellierung von Nutzervertrauen, der zur situativen Beurteilung der Angemessenheit verschiedener Systemaktionen verwendet werden kann. Die Integration des UTM in ein prototypisches SavER-System sowie die Evaluation dieses Prototypen ist Bestandteil von Kapitel 6.2. Jedes der drei Forschungskapitel wird mit einer kurzen Zusammenfassung der gesammelten Erkenntnisse abgeschlossen.

Im Schlusskapitel folgen eine kurze Zusammenfassung der Inhalte dieser Dissertation und eine Diskussion der durchgeführten Untersuchungen und der erarbeiteten Erkenntnisse. Außerdem wird es einen Ausblick auf mögliche weiterführende Arbeiten geben, die auf diese Dissertation aufbauen und sie erweitern könnten.

2 User Experience Ziele

In dieser Dissertation stellen die *Überzeugungskraft*, die *Nutzerakzeptanz* und das *Nutzervertrauen* nicht nur die maßgeblichen Qualitätskriterien für die User Experience beratender Empfehlungssysteme dar. Sie werden teilweise auch direkt in der Entscheidungsfindung der Systeme berücksichtigt. Aus diesem Grund wird im Folgenden ausführlich beschrieben, wie diese drei Konzepte innerhalb dieser Arbeit definiert und verstanden wurden.

2.1 Überzeugungskraft

Simons und Jones [Simons und Jones, 2011] definieren die *Überzeugungskraft* als den Teil der menschlichen Kommunikation, der dafür ausgelegt ist, das Urteilsvermögen und die Handlungen Anderer zu beeinflussen. Das Studium der Überzeugungskraft bezieht sich laut Chaiken und Kollegen [Chaiken et al., 1996] auf Variablen und Prozesse, die die Bildung und Änderung von Einstellungen regulieren. In dieser Arbeit liegt der Fokus vor allem darauf, Nutzer situativ von der Nützlichkeit empfohlener Aktivitäten und Maßnahmen überzeugen zu können.

Computerbasierte Programme und Informationssysteme, die dafür entworfen werden, ohne Zwang oder Täuschungen Einstellungen und Verhaltensweisen zu bestärken, zu ändern oder zu formen, werden laut Oinas-Kukkonnen und Harjuma [Oinas-Kukkonen und Harjumaa, 2008] als *persuasive Technologien* bezeichnet.

Der Zusatz, dass eine Überzeugung ohne Zwang und Täuschung erfolgen sollte, weist explizit auf die ethische Verantwortung bei der Entwicklung entsprechender Technologien hin. Diese Verantwortung wurde bereits 1999 von Berdichevsky und Neuenschwander [Berdichevsky und Neuenschwander, 1999] thematisiert. Das Zentrum ihres ethischen Frameworks zur Analyse und Entwicklung persuasiver Technologien bildete die goldene Regel, dass Entwickler persuasiver Technologien nie danach streben sollten, jemanden von etwas zu überzeugen, von dem sie selbst nicht überzeugt werden wollen.

Persuasive Technologien können Personen, laut Fogg [Fogg, 2002], auf unterschiedliche Art überzeugen. Sie können als Werkzeuge eingesetzt werden, die zum Beispiel die körperlichen Aktivitäten der Nutzer automatisch dokumentieren. Sie können als Medium agieren und ihren Nutzern zum Beispiel durch Simulationen Folgen bestimmter Verhaltensweisen aufzeigen. Außerdem können sie durch soziale Handlungen wie Lob und Belohnungen positives Verhalten bestärken.

Assistierende Empfehlungssysteme können als Medium angesehen werden, das seinen Nutzern Aktivitäten und Maßnahmen empfiehlt und durch Argumente die Vorteile dieser Aktivitäten und Maßnahmen hervorhebt [Tintarev und Masthoff, 2011]. Durch die Erweiterung ihrer Funktionalität, um die proaktive Ausführung einfacher Empfehlungen, könnten die Empfehlungssysteme auch als Werkzeug agieren, das die Nutzer entlastet und sie gleichzeitig durch die positiven Folgen dieser Maßnahmen auch zu eigenen Handlungen motiviert.

Ein gesundes Maß an Überzeugungskraft ist für beratende Empfehlungssysteme deswegen wichtig, da Empfehlungen mit finanziellen, zeitlichen oder körperlichen Aufwand einhergehen oder den Gewohnheiten der Nutzer widersprechen können.

Eine wichtige Grundlage für die Entwicklung überzeugender Systeme stellen verhaltenstheoretische Modellen wie *Foggs Behavior Model* [Fogg, 2009] oder das *Transtheoretische Modell* [Prochaska und Velicer, 1997] dar. Sie versuchen Auslöser und Abläufe von Verhaltensänderungen zu erklären, um diese anschließend auch gezielt fördern zu können. Das *Persuasive Systems Design Modell* stellt außerdem konkrete Hilfe für die Erstellung und Evaluation überzeugender Systeme bereit [Oinas-Kukkonen und Harjumaa, 2009].

2.1.1 Foggs Behavior Model

Foggs Behavior Model [Fogg, 2009] ist ein psychologisches Modell, das auf einfache Weise beschreibt, wie die Faktoren *Motivation* und *Fähigkeit* sowie der gezielte Einsatz von *Auslösern* das Eintreten von Verhaltensänderungen beeinflussen können.

- **Motivation** Die Motivation einer Person kann durch verschiedene Faktoren beeinflusst werden. Positive Beispiele sind die Aussicht auf Freude, Hoffnung oder soziale Akzeptanz. Negative Faktoren wie Schmerz, Angst oder Zurückweisung können ebenfalls zu einer Motivationssteigerung führen. Verhalten kann sowohl *intrinsisch* (aus einem inneren Anreiz wie Interesse, Neugierde oder Spaß heraus) als auch *extrinsisch* (ausgelöst durch externe Beweggründe und Konsequenzen wie materielle Vorteile oder soziale Bestärkung) motiviert sein [Ryan und Deci, 2000].

- **Fähigkeit** Die Fähigkeit, eine Aktion durchzuführen, hat beinahe noch einen stärkeren Einfluss auf das menschliche Verhalten als die Motivation. Fogg bezeichnet diesen Faktor auch als „Einfachheit“. Benötigt eine Aktion nur ein geringes Maß an Zeit, Kosten oder körperlicher oder geistiger Anstrengung, ist es möglich, dass auch eine wenig motivierte Person sie durchführt. Ein prominentes Beispiel ist die „1-Click“ Bestellfunktion bei Amazon[9]. Durch das einfache Abschließen einer Bestellung mit nur einem einzigen Mausklick sollen Nutzer dazu gebracht werden, Artikel zu bestellen, die sie bei einem größeren Aufwand womöglich nicht gekauft hätten.

- **Auslöser** Aktionen oder Verhaltensmuster können u.a. durch Systemnachrichten oder das Design bestimmter Interaktionselemente (z.B. prominente Warenkorb-Buttons) ausgelöst werden. Fogg erwähnt drei Arten von Auslösern. *Sparks* sind dafür geeignet, unmotivierte Nutzer zu motivieren. *Facilitators* sollen die Einfachheit einer Aktion oder eines Verhaltens verbessern bzw. hervorheben. Sind Motivation und Fähigkeit bereits gegeben, reichen einfache *Signals* zur Erinnerung an eine Aktion oder ein Verhalten aus.

[9] https://www.amazon.de/gp/help/customer/display.html?nodeId=201889620

Abbildung 2.1 zeigt das Zusammenspiel der drei Faktoren. Ein Vorschlag eines Verhaltens (Auslöser) hat die größte Chance auf Erfolg, wenn sowohl die Motivation als auch die Fähigkeit zur Durchführung einer Aktion oder eines Verhaltens groß sind. Allerdings kann auch eine starke Ausprägung eines der beiden Faktoren zur Aktivierung einer Person führen. Sparks und Facilitators können hierbei unterstützend wirken. Eine entscheidende Rolle spielt generell das Prinzip des *Kairos*, das Finden des richtigen Moments, um eine Nachricht zu überbringen bzw. ein Verhalten vorzuschlagen [Fogg, 2002]. In diesem Moment stehen die Chance am besten, dass ein Vorschlag angenommen und das Verhalten geändert wird.

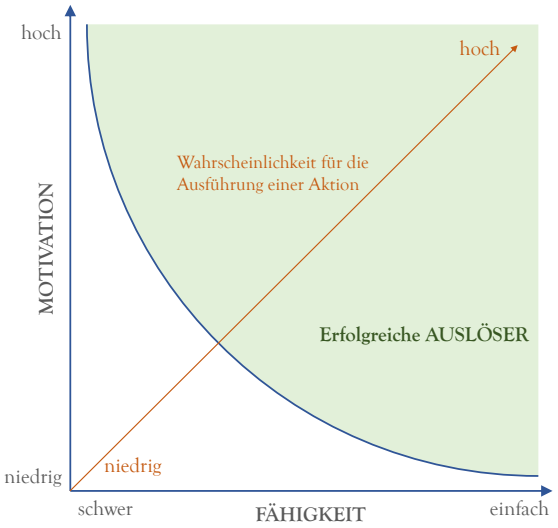

Abbildung 2.1: Fogg Behavior Model (nach [Fogg, 2009])

2.1.2 Transtheoretisches Modell

Das *Transtheoretische Modell (TTM)*[Prochaska und Velicer, 1997] geht davon aus, dass eine Verhaltensänderung kein einzelnes Ereignis ist, sondern ein länger andauernden Prozess mit diversen aufeinander folgenden Phasen. Aus diesem Grund wird es häufig auch als *Stages of Change*-Model bezeichnet. Die Phasen sind:

1. **Sorglosigkeit** Auch, wenn sich Personen in dieser Phase entweder nicht der Probleme ihres Verhaltens bewusst sind oder durch erfolglose Anläufe zur Verhaltensänderung demoralisiert sind, stellt die Sorglosigkeit eine wichtige Phase für die Förderung von Verhaltensänderungen dar. In dieser Phase ist es die Aufgabe eines persuasiven Systems, seine Nutzer auf mögliche Probleme aufmerksam zu machen oder den Anstoß für einen (weiteren) Versuch zur Verhaltensänderung zu geben.

2. **Bewusstwerdung** In dieser Phase sind sich die betroffenen Personen bereits über die Vorteile einer Verhaltensänderung im Klaren. Sie zeigen Interesse an Informationen über die Probleme ihres Verhaltens und haben die Intention, erste Veränderungen in nächster Zeit vorzunehmen. Allerdings wägen Personen

dieser Phase die möglichen Vorteile gegenüber dem Aufwand und möglichen Unannehmlichkeiten, die mit der Veränderung einhergehen, ab. Es besteht letztendlich meist noch keine Absicht, tatsächlich aktiv zu werden.

3. **Vorbereitung** Erreichen Personen diese Phase, sind sie bereit, in naher Zukunft die Verhaltensänderung anzugehen. Sie schmieden erste Pläne und loten Möglichkeiten für Verhaltensänderungen aus.

4. **Handlung** In diesem Stadium der Verhaltensänderungen werden erste während der Vorbereitung geplante Änderungen durchgeführt.

5. **Aufrechterhaltung** Personen in dieser Phase versuchen ihr neues Verhalten wiederholt auszuüben und zu manifestieren. Je länger Rückschritte in alte Verhaltensmuster vermieden werden können, umso unwahrscheinlicher werden sie. Dementsprechend werden Personen in dieser Phase auch mit der Zeit immer sicherer, dass sie das neue Verhalten beibehalten können.

6. **Anhaltende Aufrechterhaltung** Wurden die vorhergehenden Phasen erfolgreich gemeistert, besteht keinerlei Versuchung mehr, in alte Verhaltensmuster zurückzufallen. Die Personen haben das Gefühl, es selbst in der Hand zu haben, das gewünschte Verhalten aufrechtzuerhalten. Diese Phase stellt das erwünschte Optimum einer Verhaltensänderung dar, ist allerdings nur selten wirklich erreichbar. Realistischer ist es, dass Personen über einen sehr langen Zeitraum in der vorherigen Phase verweilen und sich um die Aufrechterhaltung des neuen Verhaltens bemühen.

Zwischen den einzelnen Phasen kann es zu Rückschritten kommen. Ein kompletter Rückschritt zur Sorglosigkeit ist aber äußerst selten. Allerdings sind Rückschritte von der Handlungsphase oder der Aufrechterhaltung zur Bewusstwerdung bzw. der Vorbereitung eines neuen Versuchs nicht ungewöhnlich [Prochaska und Velicer, 1997].

Ursprünglich wurde das TTM für die Analyse und Förderung einer gesunden Lebensweise entwickelt. Beispiele für die Anwendung des TTM waren die Reduktion bzw. Beendigung des Tabak- oder Alkoholkonsums oder die Förderung einer besseren Ernährung oder sportlicher Aktivitäten. Ein Überblick verschiedenster Anwendungsszenarien ist in [Prochaska, 2013] zu finden. Neben anderen übertrugen He und Kollegen [He et al., 2010] das TTM allerdings auch schon auf das Design von Systemen, die durch gezieltes Feedback nachhaltiges Verhalten fördern sollten.

Gerade am Beispiel der Förderung energiesparenden Verhaltens zeigt sich allerdings, dass das TTM auch Schwächen hat. So könnte eine Person zum Beispiel bereits erfolgreich stromsparendes Verhalten manifestiert haben, während sie sich hinsichtlich der Nutzung der Heizung oder von Fortbewegungsmitteln noch in einer früheren Phase des TTM befindet. Außerdem können Verhaltensänderungen, anstatt Schritt für Schritt zu geschehen, auch spontan durch bestimmte Ereignisse oder Druck von außen ausgelöst werden [Littell und Girvin, 2002].

Trotz dieser Schwächen stellt das TTM eine wichtige Grundlage für beratende Empfehlungssysteme dar. Eine Anpassung der Auswahl und Generierung von Empfehlungen an die Phase, in der die Zielperson sich gerade befindet, könnte die Wahrscheinlichkeit der Befolgung einer Empfehlung steigern. Die Empfehlung einfacherer Aktionen könnten in den frühen Phasen das Tätigwerden fördern (vgl. Foggs Behavior Modell). Die Empfehlung vielfältiger und neuer Aktionen könnte stattdessen die Manifestation des energiesparenden Verhaltens unterstützen. Ebenfalls in den frühen Phasen könnten personalisierte Erklärungen und Argumente Probleme im Verhalten der Nutzer verdeutlichen und wichtige Informationen für Verhaltensänderungen bereitstellen.

Neben den Phasen der Änderung beschreiben Prochaska und Velicer [Prochaska und Velicer, 1997] in ihrer Arbeit auch zehn sog. *Processes of Change*, durch die typischerweise Wechsel zwischen den Phasen ausgelöst werden. Diese Prozesse stellen Ansätze für externe Maßnahmen zur Förderung der Personen dar. Durch personalisierte Empfehlungen können zum Beispiel Verhaltensweisen vorschlagen werden, die das alte Verhalten Schritt für Schritt ersetzen sollen (*Gegenkonditionierung*). Enthält der Empfehlungstext Erklärungen und Argumente für die Durchführung der empfohlenen Maßnahmen, könnte zum einen auf Missstände aufmerksam gemacht werden (*Steigerung des Bewusstseins*). Zum anderen könnten diese Erklärungen aber auch die Auswirkungen der Empfehlungen auf die Zielperson oder ihr Umfeld unterstreichen (*Neubewertung bzgl. sich selbst oder das Umfeld*). Durch das Konzept eines proaktiven Empfehlungssystems, das die Nutzer in ihrem Alltag regelmäßig auf mögliche Maßnahmen hinweist, und durch eine vertrauensvolle und höfliche Interaktion mit den Nutzern könnten außerdem ein *unterstützendes Verhältnis* entstehen.

2.1.3 Persuasive Systems Design

Das *Persuasive Systems Design-Modell* [Oinas-Kukkonen und Harjumaa, 2009] beschreibt drei Phasen, die bei der Entwicklung persuasiver Systeme durchlaufen werden sollten, siehe Abbildung 2.2.

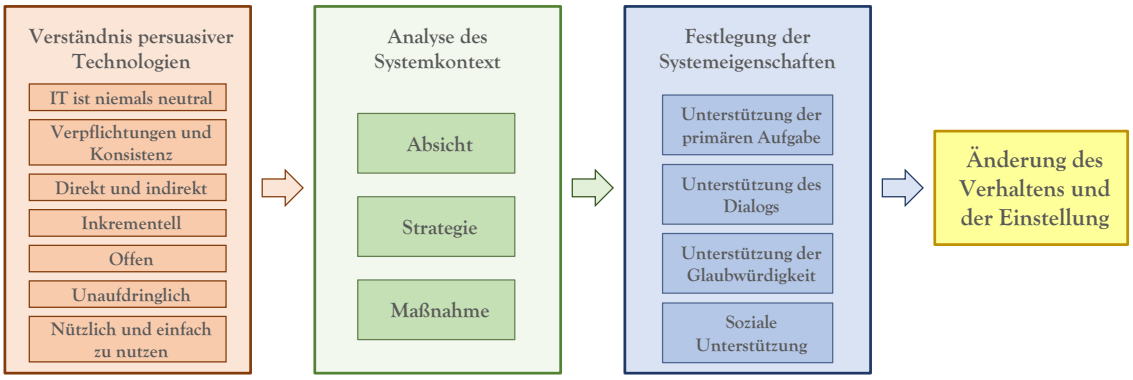

Abbildung 2.2: Phasen der Entwicklung persuasiver Systeme nach dem Persuasive Systems Design-Modell [Oinas-Kukkonen und Harjumaa, 2009]

Verständnis persuasiver Technologien Oinas-Kukkonen und Harjumaa definierten sieben sog. *Postulate* persuasiver Systeme, die typische Eigenschaften der Nutzer, grundlegende Strategien zur Überzeugung und typische Eigenschaften persuasiver Systeme beschreiben:

1. **Informationstechnologie ist niemals neutral** Sobald Menschen mit Informationstechnologie interagieren werden ihre Meinungen und ihr Verhalten in irgendeiner Form beeinflusst. Überzeugung findet demzufolge andauernd und über mehrere Schritte hinweg statt. Da die Überzeugung u.a. von den Zielen der Nutzer abhängt und diese sich während des gesamten Prozesses der Überzeugung auch ändern können, müssen sich persuasive Technologien an die geänderten Meinungen und Zielvorstellungen ihrer Nutzer anpassen, um sie Schritt für Schritt von einem neuen Verhalten zu überzeugen.

2. **Menschen bevorzugen eine geordnete und konsistente Sicht der Welt** Dieses Postulat beruht auf der Annahme, dass Einstellungen von Menschen stark von Verpflichtungen und kognitiver Konsistenz beeinflusst werden [Cialdini et al., 1981]. Gehen Menschen Verpflichtungen ein (z.B. Zielsetzungen), ist die Chance größer, dass sie sich von einer Änderung ihres Verhaltens überzeugen lassen. Eine Veränderung der Einstellung oder des Verhaltens ist ebenfalls wahrscheinlicher, wenn es gelingt, die Zielpersonen auf Inkonsistenzen bzgl. ihrer Einstellung und/oder ihres Verhaltens aufmerksam zu machen. Diese Inkonsistenzen können entweder direkt zwischen den Einstellungen und dem Verhalten bestehen oder zwischen der eigenen Einstellung und der Einstellung anderer Personen. Inkonsistenzen dieser Art werden als unangenehm und unstimmig empfunden. Ist das Gefühl der Unstimmigkeit stark genug, kann es dazu kommen, dass die Zielperson sich anpasst, um die kognitive Konsistenz wiederherzustellen.

3. **Direkter und indirekter Weg der Überzeugung** Dieses Postulat unterscheidet, ob der Inhalt einer Nachricht, die überzeugen soll, von den Empfängern im Detail beurteilt wird (direkte Überzeugung) oder ob sie die Nachricht nur flüchtig wahrnehmen und basierend auf ihren Erfahrungen oder ihrer Art darauf reagieren (indirekte Überzeugung). Der persönliche Hintergrund einer Person (z.B. Motivation, Fähigkeiten) sowie die aktuelle Situation (z.B. Stress, Beschäftigung) können beeinflussen, wie die präsentierte Information aufgenommen und verarbeitet wird.

4. **Überzeugung erfolgt schrittweise** Durch die schrittweise Präsentation von Vorschlägen und Empfehlungen ist es einfacher, Personen dazu zu bringen, eine Serie von Handlungen durchzuführen, als durch einen einmaligen zusammengefassten Vorschlag. Allerdings sollte von Beginn an das übergeordnete Ziel klar formuliert werden. Die Entscheidung für dieses Ziel sollte von der Zielperson sofort getroffen und nicht verschoben werden.

5. **Überzeugungen durch persuasive Systeme sollten immer offen sein** Um Missverständnisse zu vermeiden und die Nutzer nicht in die falsche Richtung zu lenken, sollte den Nutzern eines persuasiven Systems klar gemacht werden, dass die Entwickler des Systems bzgl. der jeweiligen Domäne voreingenommen sein könnten und Informationen manchmal auch nicht vertrauenswürdig oder gar falsch sein könnten.

6. **Persuasive Systeme sollten unaufdringlich handeln** Die Nutzer sollten während der Durchführung ihrer eigentlichen Aufgaben, bei denen sie das System unterstützt, nicht durch das System unterbrochen werden. Ist das System nicht dazu fähig, geeignete und ungeeignete Zeitpunkte für Interventionen zu unterscheiden, kann dies zu unerwünschten Resultaten führen.

7. **Persuasive Systeme sollten nützlich und einfach zu nutzen sein** Dieses Postulat fasst allgemeine Qualitätskriterien von Technologien und Systemen zusammen, die neben der Überzeugungskraft eines persuasiven Systems auch die Nutzerakzeptanz, siehe nächstes Unterkapitel, beeinflussen. Erfüllt ein System gewisse Qualitätskriterien wie Fehlerfreiheit, qualitativ hochwertige Informationen, Attraktivität oder allgemein eine positive User Experience nicht, ist es unwahrscheinlich, dass es als überzeugend wahrgenommen wird.

Analyse des Systemkontextes Mit Hilfe der erarbeiteten Grundlagen muss als nächstes der Kontext des Systems analysiert werden. Es muss beurteilt werden, welche Rolle der Überredende, der zu Überredende, die präsentierte Nachricht, der genutzte Kanal und der umgebende Kontext während der Interaktion zwischen System und Nutzer spielen. Dies dient dazu, sich über die *Ziele bzw. Absichten* des Systems sowie die Situation während der überzeugenden *Maßnahme* und geeignete *Strategien* klar zu werden.

Hinsichtlich der Absicht muss klargestellt werden, von wem der Versuch zur Überredung zu einer Änderung ausgeht und welche Änderung angestrebt wird. Der Überzeugende kann sowohl der Entwickler der persuasiven Technologie als auch die Person sein, die den Zugang zu der Technologie ermöglicht. Auch die Person, die sich ändern möchte und deswegen die Technologie nutzt, kann als Überredender fungieren. Änderungen können sowohl die Einstellung als auch das Verhalten der Zielperson betreffen. Verhaltensänderungen sind allerdings einfacher zu erreichen, als Änderungen der Einstellung.

Den Kontext des Systems bestimmen die spezifischen Voraussetzungen des Anwendungsszenarios, die individuelle Zielperson (u.a. Demographie, Persönlichkeit, Interessen, Fähigkeiten, Bedürfnisse, Ziele, bisheriges Verhalten) und die Eigenschaften der eingesetzten Technologie sowie ihre Stärken und Schwächen.

Die Strategien beschreiben letztendlich formelle und inhaltliche Faktoren bei der Vermittlung einer Verhaltensänderung sowie die Art und Weise wie die Überzeugung vonstattengehen soll (z.B. direkt oder indirekt).

Festlegung der Systemeigenschaften Bei der Entwicklung eines neuen Systems werden abschließend die Systemeigenschaften festgelegt. Bei bereits bestehenden Systemen werden die Systemeigenschaften zunächst evaluiert und dann gegebenenfalls geändert. Diesen Schritt unterstützen Oinas-Kukkonen und Harjumaa durch zahlreiche Designprinzipien, die u.a. auch auf Foggs Arbeit [Fogg, 2002] zurückgreifen. Sie befassen sich mit der Unterstützung bei der Durchführung der primären Aufgabe der Nutzer, dem Dialog zwischen System und Nutzer, der Glaubwürdigkeit des Systems sowie sozialen Faktoren. Eine komplette Liste aller Designprinzipien ist in [Oinas-Kukkonen und Harjumaa, 2009] zu finden. Für Empfehlungssysteme sind u.a. die folgenden Prinzipien von Interesse:

- **Tunneling** (Unterstützung der primären Aufgabe): Systeme haben die Möglichkeit, die Nutzer dadurch von einer Verhaltensänderung zu überzeugen, dass sie sie schrittweise durch den Prozess der Änderung geleiten. Empfehlungssysteme können dies durch die Empfehlung einzelner Maßnahmen bewerkstelligen.

- **Personalisierung** (Unterstützung der primären Aufgabe): Die Überzeugungskraft eines persuasiven Systems kann dadurch gesteigert werden, dass die bereitgestellten Informationen und Dienste an die jeweiligen Nutzer angepasst werden. In assistierenden Empfehlungssystemen sollten sowohl die ausgewählten Empfehlungen als auch die Empfehlungstexte personalisiert werden.

- **Vorschläge** (Dialog zwischen System und Nutzer): Persuasive Systeme sollten ihre Nutzer durch Vorschläge zur Ausführung von Handlungen ihres Zielverhaltens motivieren. Speziell proaktive Empfehlungssysteme können diese Aufgabe sehr gut übernehmen, da sie auch in Situationen Empfehlungen aussprechen können, in denen die Nutzer selbst nicht aktiv danach suchen.

- **Ähnlichkeit** (Dialog zwischen System und Nutzer): Systeme können überzeugender sein, wenn sie die Nutzer in gewisser Weise imitieren und somit ein Gefühl von Ähnlichkeit hervorrufen. In dieser Arbeit wurde untersucht, ob die Generierung von Empfehlungstexten basierend auf den individuellen Persönlichkeitsmerkmalen der Zielperson zu einer gesteigerten Überzeugungskraft führen kann, siehe Kapitel 5.3.

- **Vertrauenswürdigkeit** (Glaubwürdigkeit des Systems): Systeme, die als vertrauenswürdig wahrgenommen werden, besitzen eine stärkere Überzeugungskraft, als weniger vertrauenswürdige Systeme. Die Vertrauenswürdigkeit von beratenden Empfehlungssystemen könnte durch personalisierte Empfehlungstexte, siehe Kapitel 5, aber auch durch transparentes, komfortables und kontrollierbares Systemverhalten, siehe Kapitel 6, verbessert werden.

- **Expertise** (Glaubwürdigkeit des Systems): Kompetente Systeme gelten als überzeugender. Für Empfehlungssysteme bedeutet dies, dass die Auswahl der Empfehlungen qualitativ hochwertig sein muss und die Kompetenz der Systeme durch fundierte Erklärungen bewiesen werden sollte.

Prinzipien für soziale Unterstützung betreffen hauptsächlich die Berücksichtigung des Verhaltens und der Fortschritte anderer Personen. Ein Gefühl von *Kooperation* könnte in beratenden Empfehlungssystemen zum Beispiel durch den Einsatz sozialer Roboter als Interaktionsgerät oder durch geeignete Formulierungen (z.B. „Wir sollten mal wieder...“) hervorgerufen werden.

2.2 Nutzerakzeptanz

Die Forschung im Bereich *Akzeptanz von Technologien* geht bis in die 70er Jahre zurück und untersucht unter welchen Umständen Menschen Nutzerschnittstellen, Systeme und Technologien akzeptieren oder ablehnen. Dillon und Morris [Dillon und Morris, 1996] definierten die Nutzerakzeptanz als die nachweisliche Bereitschaft innerhalb einer Nutzergruppe Informationstechnologie für die Aufgaben einzusetzen, für deren Unterstützung sie entwickelt wurde.

Technology Acceptance Model Durch die steigende Zahl computerisierter Systeme wurde die Nutzerakzeptanz ein immer wichtigeres Thema. Ein prominentes Modell für Nutzerakzeptanz ist das *Technology Acceptance Model (TAM)* [Davis, 1989]. Es wurde mit dem Fokus auf in Unternehmen eingesetzte Technologien entwickelt. Dank seiner Einfachheit, aber auch seiner hohen Reliabilität wurde es allerdings auch von vielen Arbeiten in anderen Domänen genutzt, spezifiziert und erweitert. Ein recht aktueller Literaturüberblick wurde von Marangunić und Granić erstellt [Marangunić und Granić, 2015].

Der Kern des TAM blieb über die Zeit allerdings immer erhalten [Davis, 1989]. Es besagt, dass die Einstellung von Menschen gegenüber einer Technologie bzw. ihre Absicht diese Technologie zu nutzen durch zwei Hauptfaktoren beeinflusst wird, siehe Abbildung 2.3. Der *wahrgenommene Nutzen* des Systems bezieht sich auf den Grad, zu dem Nutzer denken, dass sie durch das System beim Erreichen ihrer Ziele unterstützt werden. Die *wahrgenommene Leichtigkeit der Nutzung* einer Technologie hängt zunächst einmal stark mit der Usability zusammen. Je besser alle Funktionen erkannt, gefunden und ausgeführt werden können und je leichter präsentierte Informationen verstanden werden können, desto größer ist die Chance, dass die Akzeptanz der Nutzer hoch ausfällt. Doch auch die Freude an der Nutzung hat einen starken Einfluss auf die Wahrnehmung der Nutzbarkeit des Systems [Heijden, 2004]. Sowohl der wahrgenommene Nutzen als auch die wahrgenommene Leichtigkeit der Nutzung werden durch die Eigenschaften der jeweiligen Technologie sowie durch andere externe Faktoren wie zum Beispiel Nutzereigenschaften beeinflusst [Davis, 1989]. Der Zweck der jeweiligen Technologie kann außerdem den Ausschlag dafür geben, welcher der beiden Faktoren stärker ins Gewicht fällt. Während bei arbeits- oder aufgabenorientierten Technologien mit Fokus auf Produktivität der wahrgenommene Nutzen im Vordergrund steht, spielen bei hedonisch orientierten Technologien, deren Ziel die Unterhaltung und Freude der Nutzer ist, das wahrgenommene Vergnügen und eine einfache Nutzung eine größere Rolle [Heijden, 2004].

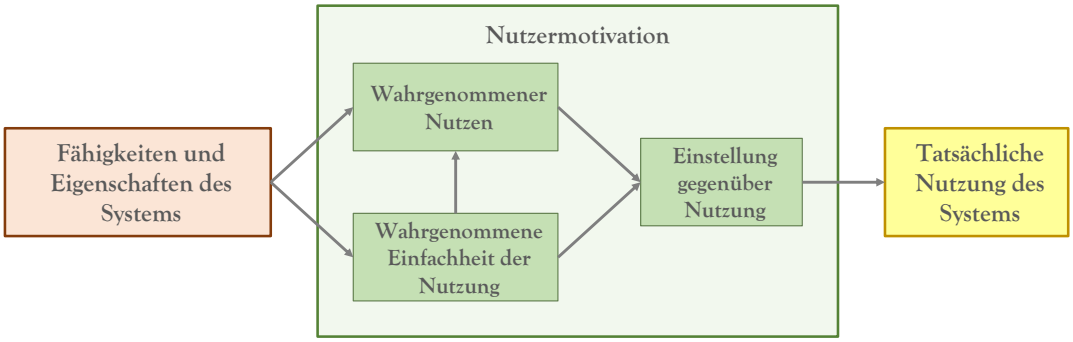

Abbildung 2.3: Technology Acceptance Model (nach [Davis, 1989])

Technology Acceptance Model 2 Im *Technology Acceptance Model 2 (TAM2)* [Venkatesh und Davis, 2000] nahmen Venkatesh und Davis zwei grundlegende Anpassungen am ursprünglichen Modell vor. Zum einen wurde die *Einstellung gegenüber der Nutzung* durch die *Intention zur Nutzung* ersetzt. Der Grund hierfür war, dass die Intention zur Nutzung eines Systems nicht zwingend von der Einstellung der Nutzer gegenüber dem System abhängt. Laut dem TAM2 gibt es einige externe Faktoren, die die wahrgenommene Nützlichkeit eines Systems und damit die Absicht einer Nutzung maßgeblich beeinflussen können, siehe Abbildung 2.4. Dazu zählen soziale Aspekte wie der Einfluss der Meinungen Anderer auf die Entscheidung der Nutzer (*Subjektive Normen*), das *Ansehen*, dass sich Nutzer durch die Verwendung des jeweiligen Systems erhoffen, und die *Freiwilligkeit* der Nutzung. Ein mildernder Faktor für die Beeinflussung durch die Meinung Anderer ist die eigene *Erfahrung*. Je größer die Erfahrung einer Person ist, umso weniger, lässt sie sich von außen beeinflussen. Auch kognitive Faktoren wirken auf die Nutzermotivation ein. Je stärker ein System bei der Erledigung von Aufgaben helfen kann (*Relevanz für Arbeit*), je besser die Ergebnisse des Systems sind (*Qualität der Ergebnisse*) und je deutlicher die erreichten Verbesserungen dem System zugeschrieben werden können (*Klarheit der Ergebnisse*), desto größer wird der Nutzen des Systems wahrgenommen.

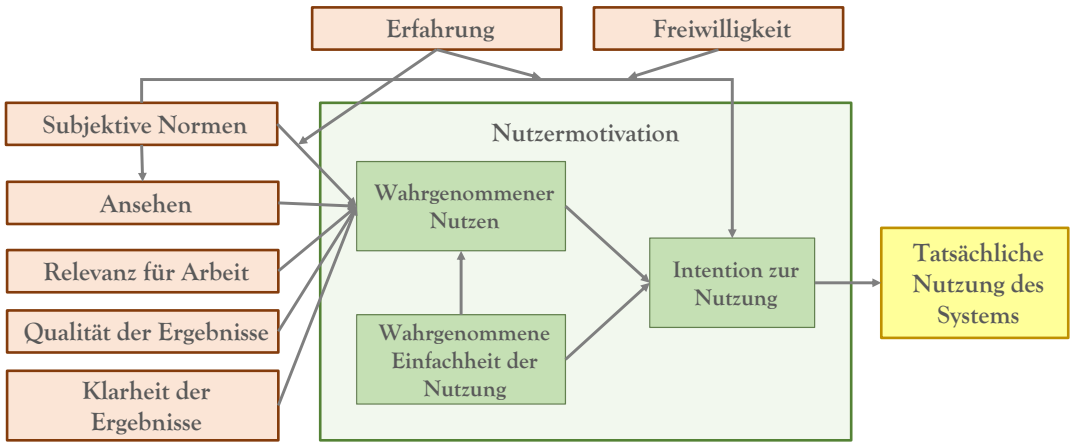

Abbildung 2.4: Technology Acceptance Model 2 (nach [Venkatesh und Davis, 2000])

Erweiterungen des TAM In einer weiteren Überarbeitung des TAM, dem *Unified Theory of Acceptance and Use of Technology (UTAUT)* ergänzten Venkatesh und Kollegen das *Geschlecht* und *Alter* der Nutzer als regulierende Faktoren der externen Einflüsse [Venkatesh et al., 2003]. Auch einen auf dem TAM basierenden Fragebogen zur Evaluation der Nutzerakzeptanz haben Davis und Venkatesh entwickelt [Davis und Venkatesh, 2004]. Dieser kann bereits in den frühen Designphasen des nutzerzentrierten Designprozesses eingesetzt werden. Um auch Begründungen für die abgegebenen Bewertungen der Nutzer zu erhalten, müssen allerdings entsprechende Fragen ergänzt werden.

TAM und Empfehlungssysteme Der Einfluss des Alters auf die Nutzerakzeptanz ist speziell im Hinblick auf ältere Menschen und das CARE-System interessant. Das Alter der Nutzer war bereits Gegenstand vieler Forschungsarbeiten (z.B. [Charness und Boot, 2009, Wagner et al., 2010]). Vor allem der Einsatz assistiver und sozialer Roboter erfreut sich großer Beliebtheit [Flandorfer, 2012, Heerink, 2011]. Das Nachlassen kognitiver Fähigkeiten wie räumliches Denken, das Schließen von Schlussfolgerungen, schnelles Denken und die Erinnerungsfähigkeit sind häufig ein Grund, dass ältere Menschen den Nutzen und die einfache Nutzung neuer Technologien nicht mehr erkennen und sie deswegen häufig ablehnen. Auch emotionale Gründe wie Unbehagen gegenüber neuen Technologien und die Angst etwas falsch zu machen sind häufig eine Ursache [Arning und Ziefle, 2007]. Diese Bedenken und Ängste gilt es zu beseitigen, während gleichzeitig die Nützlichkeit und die Einfachheit der Bedienung hervorgehoben werden müssen.

Als ein weiterer wichtiger Faktor, der starken Einfluss auf die Nutzerakzeptanz hat, stellte sich in vielen Arbeiten das Nutzervertrauen heraus (z.B. [Gefen et al., 2003, Kaasinen, 2005, Wu et al., 2011]). Bezüglich Empfehlungssystemen sind u.a. die Arbeiten von Bader und Kollegen [Bader et al., 2011b], Cramer und Kollegen [Cramer et al., 2008] und Wang und Kollegen [Wang und Benbasat, 2008] zu nennen. Auf das Nutzervertrauen wird in Kapitel 2.3 nochmals gesondert eingegangen werden, da es in dieser Dissertation einen hohen Stellenwert einnimmt. Vertrauen gegenüber einem System ist nämlich vor allem dann wichtig, wenn die Nützlichkeit und Relevanz von Empfehlungen oder automatischen Systemaktionen nicht direkt überprüft werden können und für die Nutzer ein gewisses Risiko für finanzielle, gesundheitliche oder auch nur zeitliche Nachteile besteht [Pavlou, 2003].

Dass das TAM grundsätzlich auch für Empfehlungssysteme anwendbar ist, zeigte u.a. die Forschergruppe um Pu. Studien mit Empfehlungssystemen für Musik [Jones und Pu, 2008] und Filme [Hu und Pu, 2009] zeigten, dass personalisierte und vielseitige Empfehlungen, die für die Nutzer tatsächlich relevant sind und diese unterhalten und zufriedenstellen können, sowohl den wahrgenommenen Nutzen des Systems als auch die Nutzerakzeptanz steigern können. Des Weiteren wirkten sich einfache Nutzerschnittstellen und ein geringer initialer Aufwand bis zum Erhalt interessanter Empfehlungen positiv auf die Wahrnehmung der Nutzbarkeit und damit

wiederum auf die Nutzerakzeptanz aus. Der initiale Aufwand korrelierte außerdem auch mit dem wahrgenommenem Nutzen des Empfehlungssystems, da bei einem geringem Aufwand die Genauigkeit der Empfehlungsauswahl als besser eingeschätzt wurde. Hinsichtlich der subjektiven Einschätzung des initialen Aufwands kommt es allerdings weniger auf die Dauer des Prozesses an sich an, als auf den kognitiven Aufwand bzw. die Unterhaltsamkeit der Initialisierungsphase [Hu und Pu, 2009].

Weitere Untersuchungen mit Empfehlungssystemen zeigten, dass neben dem bereits angesprochenem Nutzervertrauen auch die Transparenz der Systeme von zentraler Bedeutung für den wahrgenommenen Nutzen des Systems ist [Bader et al., 2011b, Cramer et al., 2008, Herlocker et al., 2000, Sinha und Swearingen, 2002]. Die wahrgenommene Nutzbarkeit eines proaktiven Empfehlungssystems wird außerdem stark dadurch beeinflusst, ob die Empfehlungen als aufdringlich oder ablenkend wahrgenommen werden [Bader et al., 2011b].

Neben den genannten Faktoren wird für die untersuchten Systeme außerdem darauf zu achten sein, dass die Nutzer sich durch die Empfehlungen nicht bevormundet oder beschämt fühlen. Immerhin ist es die Aufgabe dieser Systeme, die Nutzer auf Missstände oder falsche Verhaltensweisen hinzuweisen und sie dabei zu unterstützen diese Schwachstellen zu beseitigen.

Abschließend sollte darauf hingewiesen werden, dass bei Empfehlungssystemen zwischen drei Arten der Nutzerakzeptanz unterschieden werden kann: Akzeptanz und Annahme der Empfehlung, Akzeptanz und Annahme des Systems und Intention zur wiederholten Nutzung des Systems bzw. Weiterempfehlung des Systems [Cramer et al., 2008, Jones und Pu, 2008]

2.3 Vertrauen

Vertrauen ist ein subjektives, vielschichtiges und schwer fassbares Konzept. Das grundlegende Verständnis von Vertrauen bezieht sich auf die Bereitschaft sich in einer Situation „verwundbar“ zu machen [Mayer et al., 1995], da man von der Verlässlichkeit einer Person oder Sache überzeugt ist, auch wenn negative Konsequenzen möglich sind [Jøsang und Presti, 2004]. In Definition 2.1 sind wichtige Eigenschaften von Vertrauen aufgelistet [Abdul-Rahman und Hailes, 1997].

In der Informatik ist Vertrauen ein populäres Konzept, um zum Beispiel das Vertrauen zwischen Systemkomponenten in Multi-Agenten Systemen [Marsh, 1992] oder verteilten bzw. ubiquitären Systemen [Denko et al., 2011, Kiefhaber et al., 2011] zu modellieren. Das Vertrauen zwischen Nutzern ist u.a. im Bereich der sozialen Medien und sozialen Netzwerke Bestandteil vieler Forschungsarbeiten [Bhuiyan et al., 2010, Sherchan et al., 2013]. Speziell die *Reputation* von Nutzern ist hier ein wichtiges Thema [Ivanov et al., 2013]. Auch in Empfehlungssystemen wurden bereits die Vertrauensbeziehungen zwischen Nutzern berücksichtigt, um die Qualität von Filteralgorithmen verbessern zu können [Massa und Avesani, 2007a, O'Donovan und Smyth, 2005]. Hierauf wird in Kapitel 4.1 nochmals etwas genauer eingegangen.

Definition 2.1: *Vertrauen...*

- *... besteht immer zwischen zwei Entitäten (z.B. Mensch-Mensch, Mensch-System, Systemkomponente - Systemkomponente).*
- *... ist subjektiv. (Falls A Vertrauen in B hat, muss dies nicht gleichzeitig auch für C gelten.)*
- *... ist nicht symmetrisch bzw. unidirektional. (Falls A Vertrauen in B hat, muss B nicht zwingend auch Vertrauen in A haben.)*
- *... ist (bedingt) transitiv. (Wenn A B vertraut und B C vertraut, dann kann A zu einem gewissen Grad ebenfalls C vertrauen.)*
- *... ist kontextabhängig. (Vertraut A B bzgl. einer Fähigkeit X, muss dieses Vertrauen nicht zwingend auch für eine andere Fähigkeit Y gelten.)*
- *... ist dynamisch und wird durch das (häufig ungewisse) Verhalten und Aktionen der anderen Entität beeinflusst.*
- *... wird dadurch beeinflusst, inwiefern ein Verhalten oder eine Aktionen mit vorher getroffenen Annahmen und Vorhersagen übereinstimmen.*
- *... wird stärker durch Aktionen beeinflusst, die größere Auswirkungen auf einen selbst haben.*

2.3.1 Vertrauen zwischen Nutzern und Systemen

Im Fokus dieser Dissertation steht das Vertrauen, das Personen während der Nutzung eines Systems gegenüber dem System aufbauen. Nach der Terminologie von Castelfranchi und Falcone [Castelfranchi und Falcone, 2010] spricht man in diesem Fall von *affektivem Vertrauen*. Es basiert stark auf der individuellen Wahrnehmung und Einschätzung des Systems durch die Nutzer und ist eine grundlegende Voraussetzung für die Akzeptanz eines Systems [Bader et al., 2011b, Cramer et al., 2008]. Sie ist vor allem dann wichtig, wenn ein komplettes Verständnis eines Systems unmöglich oder unzweckmäßig ist [Lee und See, 2004].

Da Nutzervertrauen nur schwer direkt messbar ist, gibt es bis heute nur wenige Ansätze, die Vertrauen direkt erfassen. Leichtenstern und Kollegen [Leichtenstern et al., 2011] untersuchten, ob sich fehlendes (oder vorhandenes) Nutzervertrauen gegenüber Webseiten im Blickverhalten und in der Herzrate der Nutzer widerspiegelt. Die meisten Arbeiten zielen jedoch stattdessen darauf ab, Faktoren zu identifizieren, die das Nutzervertrauen beeinflussen (z.B. [Dzindolet et al., 2003, Lee und See, 2004]). Lee und See [Lee und See, 2004] versuchten psychologische Faktoren zu ermitteln, die beeinflussen, wie stark sich eine Person auf ein Software-System verlässt. Sie basierten ihre Untersuchungen auf der Beobachtung, dass Menschen sozial auf Technologien reagieren, und fanden u.a. heraus, dass bereits die

graphische Gestaltung einer Benutzerschnittstelle einen Einfluss auf das Nutzervertrauen hat. Van Mulken und Kollegen [van Mulken et al., 1999] wiesen den Einfluss verschiedener Medien (Text, Sprache, animierter Agent) auf das Nutzervertrauen bei der Vermittlung von Hinweisen zur Navigation durch einen Suchbaum nach. Arbeiten im E-Commerce-Bereich bestätigten die Wichtigkeit des Erscheinungsbilds, der Usability und der Nützlichkeit von Interfaces für den Aufbau von Nutzervertrauen [Hampton-Sosa und Koufaris, 2005, Koufaris und Hampton-Sosa, 2002]. Glass und Kollegen untersuchten vertrauensbasierte Kriterien für adaptive und personalisierte Anwendungen [Glass et al., 2008]. Sie entwickelten allerdings kein Modell, das zur Entscheidungsfindung in solchen Anwendungen eingesetzt werden könnte. Ein solches Vertrauensmodell entwickelten Yan und Holtmanns für mobile Applikationen [Yan und Holtmanns, 2008]. Sie versuchten nicht nur, das Vertrauen der Nutzer in die Anwendungen zu modellieren, sondern es auch zu verwalten oder gar zu stärken. Allerdings wurde das Modell lediglich in Simulationen getestet.

Das Nutzervertrauen gegenüber Empfehlungssystemen wurde ebenfalls bereits untersucht. Die Personalisierung von Empfehlungen [Komiak und Benbasat, 2006], die Empfehlung bekannter Objekte [Jannach et al., 2015] und vor allem ein transparentes Systemverhalten [Sinha und Swearingen, 2002, Swearingen und Sinha, 2002, Wang und Benbasat, 2008] können sich positiv auf das Vertrauen der Nutzer auswirken. Erklärungen sind zum Beispiel ein beliebtes Mittel, um die Ungewissheit über die Qualität von Empfehlungen zu reduzieren [Gedikli et al., 2014, Pu und Chen, 2006, Tintarev und Masthoff, 2011, Zanker, 2012]. Neben der Transparenz des Empfehlungssystems und der Qualität der Empfehlungen spielt laut Nilashi und Kollegen [Nilashi et al., 2016], ähnlich wie bei anderen Systemen, auch die Qualität der Webseite bzw. allgemeiner des User Interfaces eine wichtige Rolle bei der Förderung des Nutzervertrauens, siehe Abbildung 2.5.

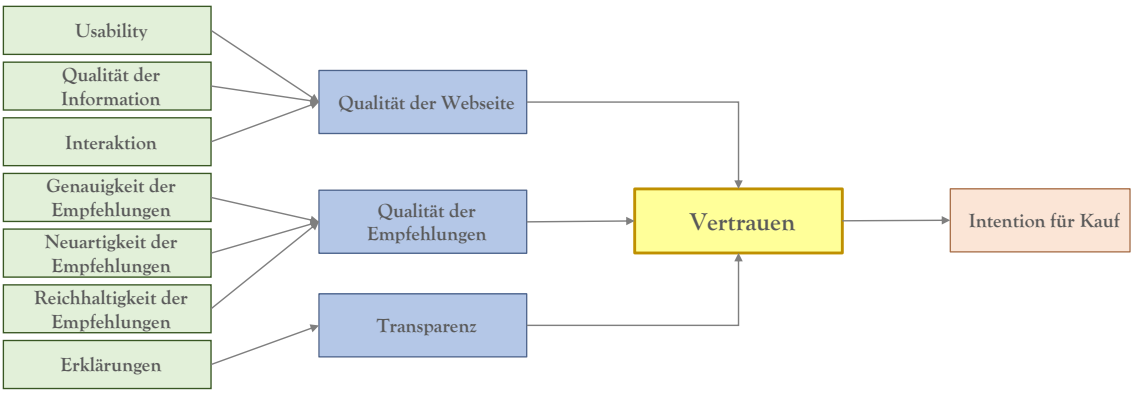

Abbildung 2.5: Modellierung von Vertrauen in E-Commerce-Empfehlungssysteme (nach [Nilashi et al., 2016])

2.3.2 Vertrauen in dieser Arbeit

Eine spezielle Herausforderung für das Nutzervertrauen gegenüber beratenden Empfehlungssystemen stellt deren Allgegenwärtigkeit und ihre Adaptivität dar. Die mitunter hohe Dynamik des Kontextes sorgt für zahlreiche Risiken, die das Nutzervertrauen gefährden können. Aus diesem Grund beeinflussen auch Faktoren das Nutzervertrauen, die offene und dynamische Computersysteme betreffen.

Um Vertrauen in eben solche Systeme modellieren zu können, wurde im Rahmen des Projekts OC-Trust versucht, Vertrauen mit Hilfe eines Sets intermediärer Dimensionen abzuleiten, sog. *Vertrauensdimensionen*. Die Dimensionen entsprechen relevanten Eigenschaften, durch die sich hoch dynamische Computersysteme auszeichnen können. Sie basieren auf einer gemeinschaftlichen Erhebung, die zu Beginn des Projekts durchgeführt wurde [Steghöfer et al., 2010]. In zusätzlichen Interviews wurden Nutzer danach befragt, welche Faktoren von Vertrauen für sie bei einer Nutzerschnittstelle von Bedeutung sind [Leichtenstern et al., 2010]. Die meisten Antworten der Teilnehmer konnten einer der folgenden Kategorien zugeordnet werden. Die in Klammern angegebenen Aussagen sind Beispiele typischer Aussagen, die von den Teilnehmern der Interviews getätigt wurden [Leichtenstern et al., 2010].

- Nutzungskomfort („Das System sollte einfach zu handhaben sein.“)
- Transparenz („Ich muss verstehen können, was das System tut.“)
- Kontrollierbarkeit („Ich möchte die Systemaktionen kontrollieren können.“)
- Privatsphäre („Das System sollte weder nach persönlichen Daten fragen, noch diese Anderen preisgeben.“)
- Zuverlässigkeit („Das System sollte stabil laufen.“)
- Sicherheit („Das System sollte Daten auf sichere Weise transportieren.“)
- Glaubwürdigkeit („Das System sollte mir von Anderen empfohlen werden.“)
- Ernsthaftigkeit („Das System sollte ein professionelles Auftreten haben.“)

Diese Vertrauensdimensionen sind auch auf assistierende Empfehlungssysteme übertragbar. *Sicherheit*, *Glaubwürdigkeit* und *Ernsthaftigkeit* decken die wichtigsten grundlegenden Eigenschaften ab, die jedes Computersystem erfüllen sollte. Die *Zuverlässigkeit* eines Empfehlungssystems korrespondiert mit der Fähigkeit, qualitativ hochwertige Empfehlungen auszuwählen. Die *Privatsphäre* spielt eine gewichtige Rolle, da es für eine bessere Personalisierung der zur Verfügung gestellten Informationen und Dienste nötig ist, mittels verschiedener am Körper getragener oder im Umfeld der Nutzer verborgener Sensoren größere Mengen privater

Daten zu sammeln. Weitere Faktoren, die das Nutzervertrauen gegenüber beratenden Empfehlungssystemen gefährden könnte, sind die hohe Heterogenität, Unsicherheit und Unvorhersehbarkeit der Situationen im Alltag der Nutzer, die zu fehlerhaften Entscheidungen des Systems und damit einer Verschlechterung der wahrgenommenen *Transparenz* und *Kontrollierbarkeit* führen können, siehe zum Beispiel [Kurdyukova et al., 2012]. Der *Nutzungskomfort* wird zu guter Letzt durch die Interaktion zwischen Nutzer und Empfehlungssystem sowie die Aufdringlichkeit des Systems beeinflusst.

3 Filtertechniken & Evaluationsmetriken

Zur Auswahl von Empfehlungen nutzen Empfehlungssysteme verschiedene *Filtertechniken*, die häufig der menschlichen Entscheidungsfindung nachempfunden sind. Über die Jahre haben sich *inhaltsbasierte*, *kollaborative* und *wissensbasierte* Filtertechniken etabliert [Jannach et al., 2010]. Da diese Filtertechniken unterschiedliche Stärken und Schwächen haben, die die Qualität der Empfehlungen beeinflussen können, wurden auch *hybride* und *kontextbewusste Filtertechniken* entwickelt, die die Schwächen der Basistechnologien durch ihre Kombination bzw. durch die Berücksichtigung von Kontextinformation beheben sollten. Die genannten Filterverfahren werden im Folgenden kurz vorgestellt. Außerdem werden Metriken präsentiert, die häufig zur Evaluation der Qualität eines Empfehlungssystems eingesetzt werden.

Insgesamt soll durch dieses Kapitel ein ausreichendes Grundwissen über Empfehlungssysteme vermittelt werden, das für das Verständnis der kommenden Kapitel und insbesondere Kapitel 4, aber auch für die Entwicklung eines eigenen (beratenden) Empfehlungssystems benötigt wird.

3.1 Inhaltsbasierte Filtertechniken

Inhaltsbasierte Ansätze zählen zu den ältesten Filtertechniken und wurden zunächst hauptsächlich für die Unterscheidung relevanter und irrelevanter Textdokumente eingesetzt [Adomavicius und Tuzhilin, 2005]. Auf den Punkt gebracht empfehlen inhaltsbasierte Systeme Objekte, die ähnlich zu Objekten sind, die einer Person in der Vergangenheit bereits gefallen haben. Auf die Meinung der Nutzer zu bereits bekannten Objekten kann aus abgegebenen Bewertungen oder indirekt aus einem Kauf oder einer Nutzung der Objekte geschlossen werden. Zum Vergleich der Objekte können Schlüsselwörter oder Bewertungen für gewisse Kategorien von Eigenschaften berücksichtigt werden. In der SavER-Anwendung könnten Energiesparempfehlungen zum Beispiel anhand des finanziellen, körperlichen oder zeitlichen Aufwands verglichen werden. Insofern die benötigten Schlüsselwörter nicht wie bei Texten automatisch extrahiert werden können, wird bei der Erstellung und Pflege der Beschreibungen Domänenwissen benötigt. Außerdem muss ein gewisser Aufwand betrieben werden, um jedes Objekt komplett und korrekt beschreiben zu können.

In dieser Dissertation wurden keine inhaltsbasierten Verfahren genutzt. In CARE sollte es zum Beispiel möglich sein, vormittags eine Sportübung, mittags ein Gericht, nachmittags ein Treffen mit anderen Leuten und abends eine Entspannungsübung zu empfehlen. In inhaltsbasierten Empfehlungssystemen können allerdings lediglich Objekte miteinander verglichen werden, die vom selben Typ sind (zum Beispiel Filme, Musik oder Texte) und durch die selben Attribute (zum Beispiel Genre, Künstler) beschrieben werden können [Jannach et al., 2010]. Außerdem neigen diese Systeme dazu, nur wenig variable Empfehlungen auszusprechen, da immer Objekte ausgewählt werden, die den bereits als „gut“ bewerteten Objekten ähnlich sind. Das *New-User-Problem* kann ebenfalls die Qualität der Empfehlungen beeinflussen. Darunter

versteht man die Problematik, dass für Nutzer, die bisher nur sehr wenige oder gar keine Bewertungen abgegeben haben, keine ausreichend durch Daten belegten Entscheidungen für oder gegen Objekte getroffen werden können.

3.2 Kollaborative Filtertechniken

Kollaborative Filtertechniken zählen zu den wichtigsten und am häufigsten verwendeten Filtertechniken. In kollaborativen Empfehlungssystemen stützt sich die Empfehlungsauswahl auf die Annahme, dass Nutzer mit ähnlichen Vorlieben für Objekte auch weiterhin ähnliche Objekte bevorzugen werden. Dementsprechend benötigt man, anders als bei inhaltsbasierten Filtern, keine Beschreibung der Objekte, um Empfehlungen auswählen zu können. Stattdessen wird das Bewertungsverhalten der Nutzer verglichen. Dadurch sind Empfehlungen auch nicht mehr nur auf eine Art von Objekten (zum Beispiel körperliche Aktivitäten, Rezepte) festgelegt. Außerdem können auch Objekte empfohlen werden, deren Eigenschaften und Attribute nur schwer in das System eingepflegt oder miteinander verglichen werden können.

Eine Variante des kollaborativen Filterns, auf die in dieser Arbeit zurückgegriffen wird, ist das *nutzerbasierte* kollaborative Filtern. Es setzt sich aus zwei Schritten zusammen: (1) Finden der k ähnlichsten Nutzer des Zielnutzers u. (2) Abschätzen einer Bewertung $\tilde{r}$ für ein bisher vom Zielnutzer nicht bewertetes Objekt i.

Finden der ähnlichsten Nutzer Es gibt unterschiedliche Ähnlichkeitsmaße, wie die mittlere quadratische Differenz oder den Kosinus [Adomavicius und Tuzhilin, 2005, Ahn, 2008]. Das Maß, das am effektivsten funktioniert und darum am häufigsten Verwendung findet, ist die Pearson-Korrelation [Breese et al., 1998, Jannach et al., 2010]. Ihr Vorteil ist, dass sie das unterschiedliche Bewertungsverwalten der Nutzer berücksichtigt. Während bei anderen Ähnlichkeitsmaßen Nutzer eine stärkere Rolle einnehmen, die die Bewertungsskala stärker ausnutzen, findet bei Pearson ein Vergleich statt, bei dem alle Nutzer in etwa gleich behandelt werden. Die Ähnlichkeit zweier Nutzer u und v berechnet sich nach folgender Gleichung:

$$sim(u,v) = \frac{\Sigma_{i \in I_u \cap I_v}(r(u,i) - \bar{r}(u))(r(v,i) - \bar{r}(v))}{\sqrt{\Sigma_{i \in I_u \cap I_v}(r(u,i) - \bar{r}(u))^2}\sqrt{\Sigma_{i \in I_u \cap I_v}(r(v,i) - \bar{r}(v))^2}} \tag{3.1}$$

$r(u,i)$ bzw. $r(v,i)$ stehen für die Bewertungen, die die zu vergleichenden Nutzer u und v für ein Objekt i abgegeben haben. $\bar{r}(u)$ und $\bar{r}(v)$ stehen für die durchschnittlichen Bewertungen. I_u und I_v beschreiben die Menge an Objekten, die die Nutzer bereits bewertet haben. $I_u \cap I_v$ ist die Menge an Objekten, die von beiden Nutzern Bewertungen erhalten haben. Für die Berechnung der Ähnlichkeit der Nutzer werden lediglich Bewertungen für Objekte aus dieser Menge berücksichtigt. Die berechnete Korrelation ist umso zuverlässiger, je mehr Objekte von beiden Nutzern bewertet wurden.

Um diesen Aspekt auch in der Abschätzung der Ähnlichkeit zu berücksichtigen, haben Linden und Kollegen [Linden et al., 2003] Pearsons Korrelationskoeffizienten um einen Faktor ergänzt, der Ähnlichkeiten, die auf einer kleinen Menge gemeinsam bewerteter Objekte beruht, abwertet.

$$sim'(u,v) = \frac{min(|I_u \cap I_v|, \gamma)}{\gamma} * sim(u,v) \tag{3.2}$$

γ ist eine vordefinierte Konstante, die verwendet wird, um den Einfluss der Mengengröße zu normalisieren. Hu und Pu wählten in ihrer Arbeit, auf die im Laufe dieses Kapitels noch näher eingegangen werden wird, zum Beispiel $\gamma = 5$ [Hu und Pu, 2011].

Nach der Berechnung aller Ähnlichkeiten werden die k ähnlichsten Nutzer ausgewählt, deren Bewertungen im nächsten Schritt verwendet werden, um eine wahrscheinliche Bewertung des Zielnutzers für ein neues Objekt abzuschätzen.

Abschätzen wahrscheinlicher Bewertungen Auch für die Aggregation der Bewertungen der ähnlichsten Benutzern (Menge S) gibt es unterschiedliche Funktionen. Einfache Varianten bilden den Durchschnitt. Andere verrechnen die Ähnlichkeit der Nutzer in einer gewichteten Summe. Damit sich die Bewertungen von Nutzern, die extremere Bewertungen abgeben, nicht zu stark in der berechneten Bewertung niederschlagen, ist es allerdings empfehlenswert, eine sog. *angepasste, gewichtete Summe* zu verwenden [Jannach et al., 2010], siehe Gleichung (3.3).

$$\tilde{r}(u,i) = \bar{r}(u) + \frac{\Sigma_{v \in S} sim(u,v) * (r(v,i) - \bar{r}(v))}{\Sigma_{v \in S} |sim(u,v)|} \tag{3.3}$$

Eine Schwäche kollaborativer Empfehlungssysteme sind die sogenannten *Cold-Start-Probleme* [Jannach et al., 2010]. Darunter fällt das bereits erwähnte *New-User-Problem*, aber auch das *Sparsity-Problem*, das durch eine insgesamt geringe Menge an Bewertungen über alle Nutzer und Objekte hinweg ausgelöst wird. In beiden Fällen können die Nutzer entweder gar nicht oder nur sehr unzuverlässig miteinander verglichen werden, da nur wenige Überlappungen bei den bereits bewerteten Objekten bestehen. Eine Folge ist, dass neue Bewertungen nur ungenau vorhergesagt werden können und anschließend eventuell unpassende Empfehlungen ausgesprochen werden. Dadurch können (vor allem neue) Nutzer kein Vertrauen in das System aufbauen und werden es womöglich nicht mehr länger nutzen.

3.3 Wissensbasierte Filtertechniken

Wissensbasierte Filtertechniken sind vor allem für Systeme geeignet, in denen Expertenwissen (zum Beispiel von Ernährungsberatern oder medizinischem Personal) in die Empfehlungsauswahl miteinbezogen werden soll. Anders als inhaltsbasierte und kollaborative Systeme sind wissensbasierte Systeme nicht auf Nutzerbewertungen angewiesen und somit auch nicht anfällig für das Cold-Start-Problem. Sie nutzen

domänenspezifisches Wissen, um die Objekteigenschaften mit den Nutzeranforderungen zu vergleichen, um anschließend die Objekte auszuwählen zu können, die die Anforderungen am besten erfüllen. [Burke, 2000]

Die Empfehlungsauswahl kann auf verschiedene Arten erfolgen. Bei der *regelbasierten Auswahl* werden basierend auf den Nutzeranforderungen Regeln definiert, die die Objekte erfüllen müssen [Jannach et al., 2010]. Bei der *fallbasierten Auswahl* werden dagegen die Ähnlichkeiten zwischen den aktuellen Nutzeranforderungen und den Nutzeranforderungen in früheren Fällen mit erfolgreichen Empfehlungen berechnet und diese Empfehlungen wiederverwendet [Jannach et al., 2010]. Auch auf *Ontologien basierende Verfahren* wären denkbar [Middleton et al., 2009]. Sie bergen den Vorteil, dass für Themenbereiche wie zum Beispiel Medizin, Ernährung und körperliche Aktivitäten bereits eine Vielzahl an Ontologien mit modelliertem Fachwissen verfügbar ist. Darüber hinaus eignen sich Ontologien generell sehr gut zur Modellierung von Nutzerdaten und Kontextinformationen. Allerdings ist zu beachten, dass für die Anpassung oder Erstellung von Ontologien umfangreiches Domänenwissen benötigt wird. Außerdem geht mit den Prozessen zum Ziehen von Schlussfolgerungen ein hoher Berechnungsaufwand einher, der die Verwendung von Ontologien für die Anwendungsbeispiele in dieser Arbeit unattraktiv macht.

Da die Qualität der Empfehlungen stark von der Menge und Qualität der Nutzeranforderungen abhängt, ist es bei regelbasierten sowie bei fallbasierten Systemen häufig möglich bzw. sogar nötig, dass die Nutzer ihre Anforderungen iterativ anpassen und präzisieren. Bei unzureichenden Anforderungen können womöglich zu wenige Objekte ausgefiltert werden, um eine eindeutige Empfehlung aussprechen zu können. Bei zu spezifischen oder sich widersprechenden Anforderungen kann es wiederum sein, dass keines der Objekte den Anforderungen entspricht und keine Empfehlung mehr möglich ist. Da sich der Auswahlprozess wissensbasierter Systeme allerdings anhand ausschlaggebender Anforderungen gut für die Nutzer beschreiben lässt, können diese sowohl bei der Anpassung der Anforderungen als auch generell bei der Entscheidungsfindung unterstützt werden. [Burke, 2000]

Aufgrund der genannten Eigenschaften werden wissensbasierte Verfahren häufig in Systemen eingesetzt, in denen Nutzer seltener Bewertungen abgeben (z.B. teure Elektrogeräte, Immobilien, Restaurants). Durch den Einsatz der Verfahren in beratenden Empfehlungssystemen könnten spezielle Nutzeranforderungen (z.B. „Ich benötige Aktivitäten, die bei Arthrose im Knie helfen.“), aber auch spezielle kontextuelle Situationen (z.B. „Ich möchte Outdoor-Aktivitäten wie Fahrrad fahren nur empfohlen bekommen, wenn es mindestens 10°C warm und sonnig oder heiter ist.“) berücksichtigt werden.

3.3.1 Regelbasierte Auswahl

Für die regelbasierte Empfehlungsgenerierung werden Attribute benötigt, mit denen die Nutzeranforderungen und Objekte beschrieben werden können. Durch geeignete Operatoren wie zum Beispiel =, <, >, <=, >=, *in* können die Anforderungen der

Nutzer für die einzelnen Attribute durch Regeln formuliert werden. Bei der Empfehlungsgenerierung werden dann alle Objekte gesucht, die die Anforderungen erfüllen. Zur Umsetzung komplexer Regelsysteme kann auf sog. *Rule Engines* (z.B. JRuleEngine[10]) zurückgegriffen werden.

Schematisch dargestellte Beispiele für eine regelbasierte Empfehlungsauswahl sind in Abbildung 3.1 zu sehen. Die Beispiele zeigen auch, dass sowohl die Eigenschaften der Objekte als auch der situative Kontext für die Filterung der Empfehlungen berücksichtigt werden können. Für das untere Beispiel ist anzumerken, dass die Prüfung alle Objekte der Kategorie „Outdoor-Aktivität“ betrifft.

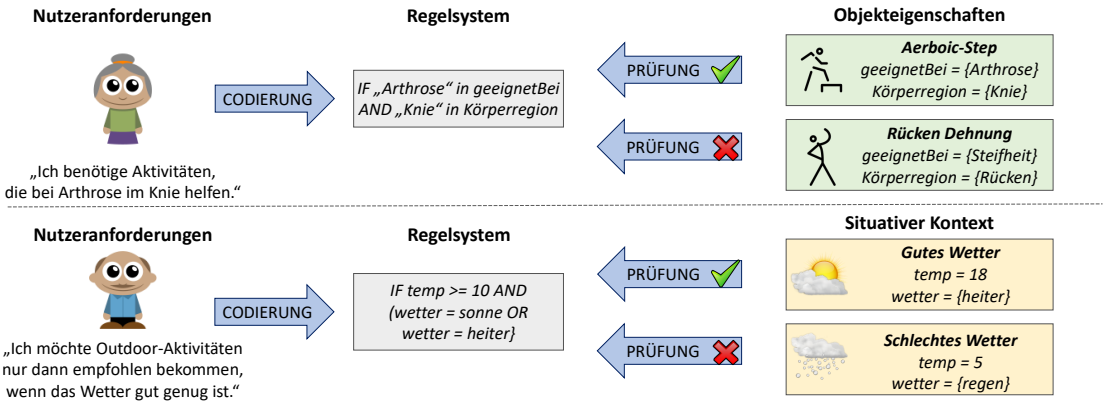

Abbildung 3.1: Schematischer Ablauf einer regelbasierten Empfehlungsauswahl

Analog zur regelbasierten Filterung ist auch eine Filterung mittels Constraints oder Datenbank-Anfragen möglich [Jannach et al., 2010].

3.3.2 Fallbasierte Auswahl

Anders als bei der regelbasierten Auswahl berücksichtigt die fallbasierte Auswahl von Empfehlungen ehemalige Fälle, um in der aktuellen Situation die besten Empfehlungen zu finden. Man orientiert sich dabei an den Phasen des fallbasierten Schließens [Aamodt und Plaza, 1994], siehe Abbildung 3.2. Zunächst wird die Ähnlichkeit zwischen dem aktuellen Fall und ehemaligen Fällen berechnet (Phase: Retrieve). Anschließend wird versucht, die erfolgreichen Ergebnisse der ähnlichsten Fälle wiederzuverwenden (Phase: Reuse). Sollten diese Ergebnisse nicht direkt auf den aktuellen Fall angewendet werden können, ist eine Anpassung nötig (Phase: Revise). Hier unterscheidet sich das Vorgehen fallbasierter Empfehlungssysteme vom fallbasierten Schließen. Beim fallbasierten Schließen werden die gefundenen Lösungen an das aktuelle Problem angepasst. Dies ist mit Objekten in einem Empfehlungssystem meistens allerdings nicht möglich. Hier wird versucht, die Nutzeranforderungen in einem weiteren Schritt so anzupassen, dass die nächste Iteration eine bessere Empfehlung hervorbringt. Ein klassisches Beispiel für eine weitere Anforderung wäre zum Beispiel „ein billigeres Objekt als das vorgeschlagene Objekt“. Abschließend soll das

[10] http://jruleengine.sourceforge.net/

System aus dem aktuellen Fall lernen (Phase: Retain). Hierfür wird die Beschreibung der aktuellen Situation und das erfolgreich empfohlene Objekt gespeichert. Möchte man noch mehr Wissen aus dem aktuellen Fall gewinnen, können auch abgelehnte Objekte gespeichert werden.

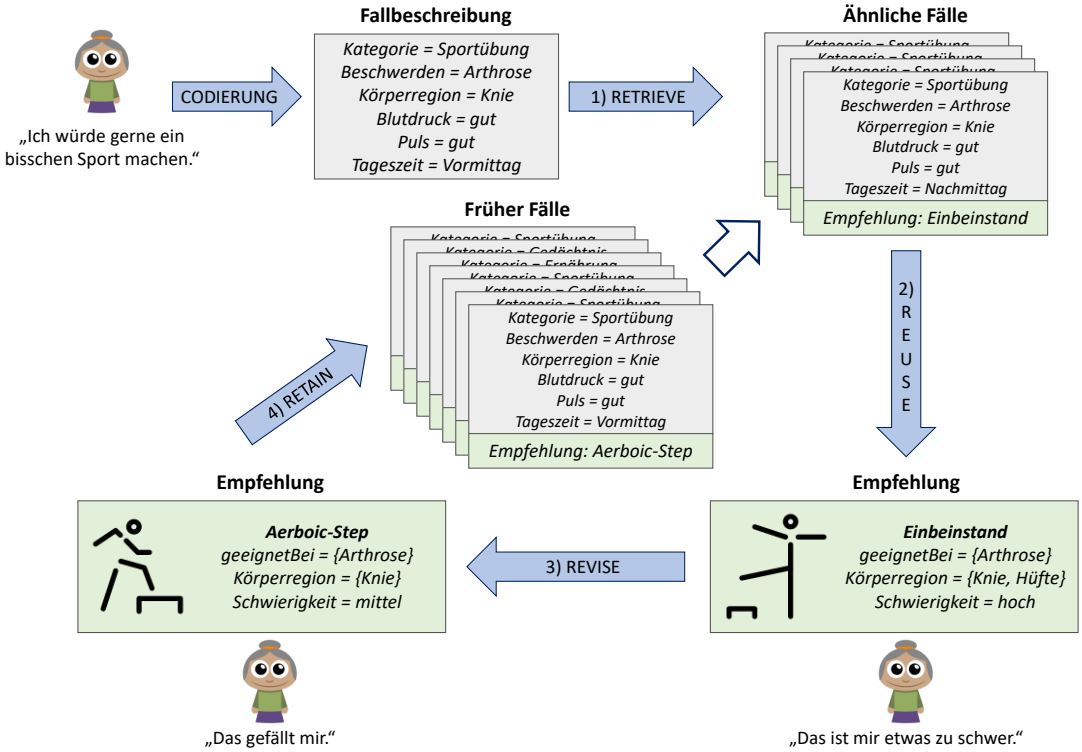

Abbildung 3.2: Schematischer Ablauf einer fallbasierten Empfehlungsauswahl

3.4 Hybride Filtertechniken

In hybriden Empfehlungssystemen werden zwei oder mehr Filterverfahren so kombiniert, dass möglichst viele der Nachteile der einzelnen Verfahren behoben und somit präzisere Empfehlungen ausgesprochen werden können. Eine ausführliche Diskussion hybrider Empfehlungssysteme ist in den Arbeiten von Adomavicius und Tuzhilin [Adomavicius und Tuzhilin, 2005] und Burke [Burke, 2002] zu finden. Im Folgenden soll nur ein kurzer Überblick gegeben werden, um das Verständnis der im Verlauf des Kapitels beschriebenen Algorithmen für CARE und SavER zu erleichtern.

Häufig werden kollaborative Filter mit Filtern kombiniert, die aufgrund ihres zusätzlichen Wissens über die Eigenschaften der Objekte weniger abhängig von Nutzerbewertungen sind. Die Kombination der Filtertechniken kann parallel, sequentiell und auch monolithisch erfolgen [Jannach et al., 2010].

Parallele Nutzung der Methoden Jede Filtertechnik wählt separat Empfehlungen aus. Die Ergebnisse werden anschließend miteinander kombiniert. Im einfachsten Fall werden alle Empfehlungen in einer Ergebnisliste präsentiert (*Mischung*).

Soll bei der Kombinierung nochmals eine Filterung erfolgen, können die errechneten Scores der Filtertechniken gewichtet aggregiert werden (*Gewichtung*). Die Gewichte können auch situativ gewählt werden, so dass immer nur die Ergebnisse der jeweils „besten“ Filtertechnik(en) berücksichtigt werden (*Umschaltung*). Was „besser“ bedeutet, hängt von der jeweils gewählten Evaluationsmetrik ab, siehe Kapitel 3.6.

Sequentielle Nutzung der Methoden Bei der sequentiellen Kombination kann eine Filtertechnik die Ergebnisse der anderen Filtertechnik verfeinern. Burke führte beispielsweise eine wissensbasierte Auswahl generell geeigneter Objekte durch und ließ anschließend nur diese Objekte durch kollaboratives Filtern bewerten und ordnen (*Kaskade*) [Burke, 2000]. Eine weitere Variante der sequentiellen Kombination ist die Erzeugung von Nutzermodellen durch eine Filtertechnik und die Verwendung der Modelle durch eine weitere (*Meta-Level*). Pazzani nutzte zum Beispiel inhaltsbasierte Nutzermodelle, die die Eigenschaften präferierter Objekte enthielten, zur Auswahl der ähnlichsten Nutzer in einem kollaborativen Filter [Pazzani, 1999].

Monolithische Kombination der Methoden Die Grundidee dieser Art der Kombination ist, das Wissen mehrerer Filtertechnologien zu kombinieren und anschließend in einem einzelnen Verfahren einzusetzen. Werden bereits vorhandenen Daten kombiniert, spricht man von *Merkmalskombination*. Um das New-User-Problem zu verringern, nutzte Pazzani [Pazzani, 1999] zum Beispiel in einem kollaborativen Empfehlungssystem neben den abgegebenen Bewertungen der Nutzer auch ihre demographischen Daten. Erzeugt eine Filtertechnik durch ihre Ergebnisse neues Wissen, das von einer anderen Filtertechnik zur Empfehlungsauswahl genutzt werden kann, ist das eine *Merkmalserweiterung*. Sarwar und Kollegen [Sarwar et al., 1998] erzeugten beispielsweise für eine kollaborative Filterung zusätzliche fiktive Nutzer (Bots), die zu empfehlende Nachrichtentexte anhand wissensbasierter Regeln (z.B. Anzahl an Rechtschreibfehlern) bewerteten.

3.5 Kontextbewusste Filtertechniken

Wie bereits erwähnt können Kontextinformationen, die die aktuelle Situation der Nutzer genauer beschreiben, dazu genutzt werden, die Empfehlungsgenerierung besser an diese Situation anzupassen. Am häufigsten werden die aktuelle Position und die Zeit (Physikalischer Kontext) verwendet. Aber auch Kontextinformationen über die Umwelt (z.B. Wetter oder Licht) oder der soziale Kontext (z.B. Personen in der Nähe der Nutzer) können abhängig vom Anwendungsfall hilfreich sein. Für die in diesem Kapitel beschriebenen Untersuchungen stand vor allem der persönliche Kontext der Benutzer im Vordergrund. Hierzu gehören u.a. die Persönlichkeit der Nutzer, ihre Stimmung, ihr Wohlbefinden oder auch ihr Verhalten, siehe Kapitel 4.1.

Davon ausgehend, dass der Prozess der Empfehlungsgenerierung aus den drei Komponenten *Dateneingabe*, *Empfehlungsauswahl* und *Ausgabe der Empfehlungsliste* besteht, beschrieben Adomavicius und Tuzhilin [Adomavicius und Tuzhilin, 2011]

drei Paradigmen zur Berücksichtigung von Kontextinformationen in diesen Komponenten. Eine Voraussetzung für die Anwendung der Paradigmen ist, dass der Kontext in einem System immer durch die selbe vordefinierte Menge von Attributen beschrieben wird und sich die Struktur dieser Informationen nicht ändert.

Kontextbasierte Vorfilterung In diesem Fall der kontextbewussten Filterung werden nur Objekte und Bewertungen für die Empfehlungsauswahl berücksichtigt, die für den aktuell vorliegenden Kontext relevant sind. Der vorliegende Kontext wird also bereits vor der eigentlichen Filterung berücksichtigt, um die Eingabedaten zu reduzieren, siehe Abbildung 3.3 (1). Möchte eine Person zum Beispiel an einem Samstag Empfehlungen für Aktivitäten erhalten, werden Aktivitäten, die nur für andere Wochentage geeignet sind, ausgefiltert und auch Bewertungen für Aktivitäten, die an einem anderen Tag durchgeführt wurden, werden ignoriert. Ein Vorteil dieses Vorgehens ist, dass nach der Vorfilterung die bekannten Filtertechniken unverändert genutzt werden können und dennoch zwischen den kontextabhängigen Präferenzen der Nutzer unterschieden werden kann. Durch die verringerte Datenmenge kann außerdem der Rechenaufwand während der eigentlichen Empfehlungsauswahl reduziert werden. Die verringerte Datenmenge kann allerdings auch von Nachteil sein, da verstärkt das Cold-Start-Problem auftreten kann. Es ist also wichtig, ein geeignetes Maß für die Granularität der Kontextinformationen zu finden. Auf das genannte Beispiel bezogen, könnte durch eine Generalisierung des Kontextes statt zwischen den einzelnen Wochentagen auch zwischen Wochenenden und Arbeitstagen unterschieden werden. Eine feinere Unterscheidung bringt in den meisten Fällen keinen signifikanten Vorteil.

Kontextbasierte Nachfilterung Bei diesem Ansatz wird der vorliegende Kontext zunächst ignoriert und eine klassische Empfehlungsgenerierung mit allen Daten durchgeführt, siehe Abbildung 3.3 (2). Erst danach wird die Ergebnisliste an die kontextuelle Situation des Nutzers angepasst und Empfehlungen, die im Bezug auf den aktuellen Kontext gewisse Eigenschaften nicht besitzen oder nur eine geringe Relevanz haben, werden ausgefiltert. Beispiele wären eine Empfehlung für ein Abendessen zum Frühstück oder das Trocknen frisch gewaschener Wäsche im Freien bei schlechtem Wetter. Etwas weniger restriktiv ist die Gewichtung und Neuordnung der Empfehlungen anhand der Anzahl passender Eigenschaften oder der Wahrscheinlichkeit der Relevanz. Der Vorteil der kontextbasierten Nachfilterung ist, dass die bekannten Filtertechniken unverändert eingesetzt werden können.

Kontextbasierte Auswahl Im Vergleich zu den anderen Methoden ist die kontextbasierte Auswahl komplexer, da die verwendete Filtertechnik so erweitert werden muss, dass sie den Kontext direkt berücksichtigen kann, siehe Abbildung 3.3 (3). Ein heuristischer Ansatz hierfür wäre, die Ähnlichkeit verschiedener kontextueller Situationen abzuschätzen, um anschließend die in den einzelnen Situationen vergebenen Bewertungen anhand dieser Ähnlichkeiten zu gewichten.

Die Bewertungen früherer Situationen, in denen der Kontext ähnlich zur aktuellen Situation ist, würden dann stärker bei der Empfehlungsauswahl berücksichtigt als die Bewertungen unähnlicher Situationen. Würde man als Gewichtungen nur 0 und 1 zulassen, würde dies wiederum einer kontextbasierten Vorfilterung entsprechen. Anstatt heuristischer Ansätze sind auch modellbasierte Verfahren möglich, die zum Beispiel Bayes'sche Netze oder Support Vector Machines verwenden. Diese Ansätze werden allerdings in dieser Arbeit nicht weiter berücksichtigt. Für eine detaillierte Beschreibung der Verfahren sei deswegen auf die Arbeit von Adomavicius und Tuzhilin [Adomavicius und Tuzhilin, 2011] verwiesen.

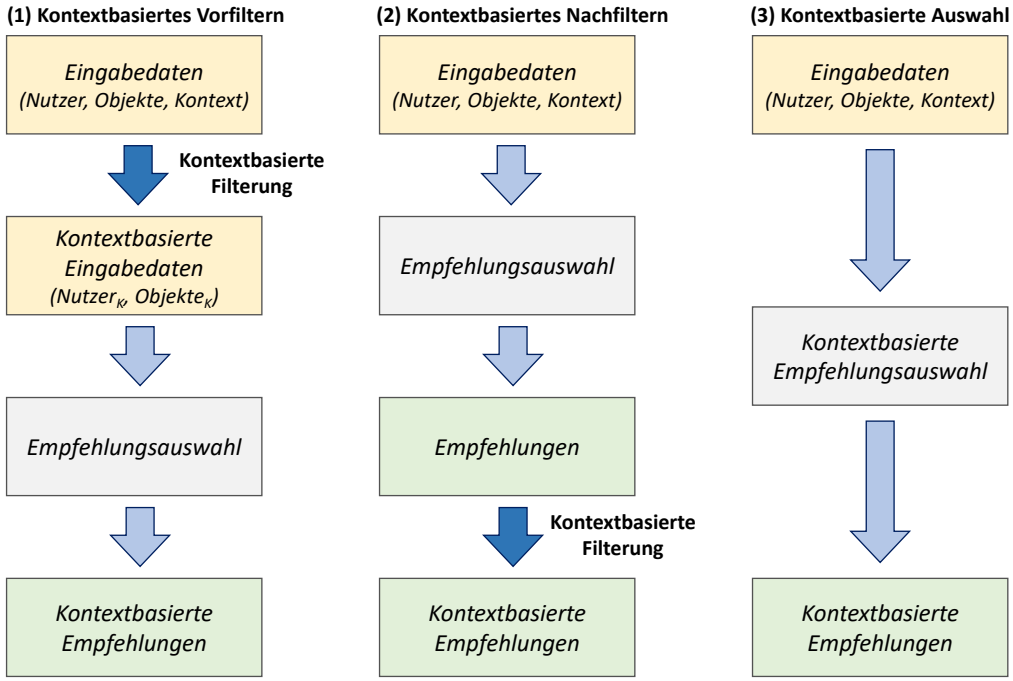

Abbildung 3.3: Paradigmen zur Berücksichtigung von Kontextinformatione bei der Empfehlungsgenerierung (nach [Adomavicius und Tuzhilin, 2011])

Welches dieser Verfahren genutzt wird, hängt stark vom jeweiligen System und dem berücksichtigten Kontext ab. Während zum Beispiel in CARE die Tageszeit oder das Wetter ein K.o.-Kriterium für bestimmte Aktivitäten darstellen, beeinflussen der soziale Kontext oder der Wochentag die Relevanz einer Aktivität nur zu einem gewissen Maß. Dementsprechend wären die Tageszeit und das Wetter gut für eine Vorfilterung geeignet und der soziale Kontext und der Wochentag eher für eine Nachfilterung mittels Gewichten oder eine kontextbasierte Auswahl. Auch eine Kombination der Verfahren ist denkbar.

3.6 Evaluationsmetriken

Um geeignete Metriken zur Evaluation eines Empfehlungssystems auswählen zu können, muss man sich zunächst darüber klar werden, anhand welcher Kriterien das

System bewertet werden soll. Viele der häufig genutzten Metriken bewerten die Genauigkeit der vorhergesagten Bewertungen. Andere Metriken beurteilen die Genauigkeit der Klassifikation der Objekte in relevante und nicht relevante Objekte und wieder andere Metriken messen die Korrektheit der Sortierung der empfohlenen Objekte hinsichtlich des Nutzerinteresses. Da in den untersuchten Anwendungsfällen allerdings keine geordneten Ergebnislisten präsentiert werden, findet die zuletzt genannte Variante in dieser Arbeit keine Berücksichtigung. Eine weitere interessante Art von Kriterien beschreibt, wie vielen der Nutzer das System „aussagekräftige" Empfehlungen präsentieren kann bzw. wie viele der möglichen Objekte tatsächlich empfohlen werden. „Aussagekräftig" bedeutet in diesem Fall, dass die vorhergesagte Bewertung für ein Objekt auf den Bewertungen anderer Nutzer und nicht nur zum Beispiel auf der durchschnittlichen Bewertung der Zielperson beruht.

3.6.1 Vorhersage von Bewertungen

Mean Absolute Error (MAE) Der mittlere absolute Fehler, wie er im Deutschen genannt wird, ist eines der am häufigsten verwendeten Maße im Bereich Information Retrieval und Empfehlungssysteme. Er berechnet die durchschnittliche Abweichung zwischen den vorhergesagten Nutzerbewertungen $\tilde{r}_i$ und den tatsächlich abgegebenen Bewertungen r_i [Herlocker et al., 2004].

$$MAE = \frac{\sum_{i=1}^{n} |\tilde{r}_i - r_i|}{n} \tag{3.4}$$

n steht in diesem Fall für die Anzahl der evaluierten Empfehlungen. Je kleiner der MAE ausfällt, umso besser ist die Qualität der vorhergesagten Bewertungen.

Um die Ergebnisse unterschiedlicher Applikationen auch trotz möglicherweise abweichender Bewertungsintervalle miteinander vergleichen zu können, entwickelten Goldberg und Kollegen [Goldberg et al., 2001] eine normalisierte Variante des MAE.

$$NMAE = \frac{MAE}{r_{max} - r_{min}} \tag{3.5}$$

Der NMAE berücksichtigt die maximale Spanne der Bewertungen, die sich durch r_{max} und r_{min} berechnen lässt, und liefert Werte im Intervall von 0 bis 1.

Root Mean Square Error (RMSE) Der RMSE beschreibt ebenfalls die Abweichung zwischen den vorhergesagten und den tatsächlichen Bewertung. Durch das Quadrieren der Differenzen fallen größere Abweichungen jedoch stärker ins Gewicht.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\tilde{r}_i - r_i)^2}{n}} \tag{3.6}$$

3.6.2 Einschätzung der Relevanz

Precision, Recall und F1 Zwei Maße, die sehr häufig bei der Evaluation von Klassifikationsergebnissen verwendet werden, sind *Precision* und *Recall*. Die Precision gibt an, wie viele der Objekte, die einer bestimmten Klasse zugeordneten wurden,

tatsächlich korrekt zugeordnet wurden. Der Recall berechnet dagegen den Anteil der korrekt einer Klasse zugeordneten Objekte an den insgesamt tatsächlich zur Klasse gehörigen Objekten. In Empfehlungssystemen wird mit Hilfe dieser Maße beurteilt, wie gut das System darin ist, vorherzusehen, ob ein Objekt für den Zielnutzer relevant ist oder nicht. „Relevanz“ ist in diesem Fall mit einer voraussichtlich hohen Bewertung gleichzusetzen. Können Objekte zum Beispiel auf einer Skala von 1 = „sehr schlecht“ bis 5 = „sehr gut“ bewertet werden, könnte man die Grenze, ab der Objekte als „relevant“ klassifiziert werden, auf 3,5 legen.

Die Precision berechnet in diesem Fall den Anteil der berechtigterweise empfohlenen Objekte ($HITS$) an den insgesamt empfohlenen Objekten (REC).

$$P_u = \frac{|HIT_u|}{|REC_u|} \tag{3.7}$$

Der Recall gibt den Anteil der berechtigterweise empfohlenen Objekte ($HITS$) an den tatsächlich relevanten Objekten (REL) an.

$$R_u = \frac{|HIT_u|}{|REL_u|} \tag{3.8}$$

Im Gegensatz zu MAE und RMSE spiegeln der Recall und vor allem die Precision stärker die Nutzererfahrung wider [McLaughlin und Herlocker, 2004]. Nutzer erhalten in den meisten Empfehlungssystemen keine Bewertungsvorhersagen, sondern entscheiden nur, ob eine Empfehlung für sie interessant ist oder nicht.

Bei einer reinen Betrachtung von Precision und Recall tritt allerdings ein Problem auf. Versucht man durch eine Erhöhung oder Reduzierung der Anzahl der empfohlenen Objekte eines der beiden Maße zu stärken, geht dies meistens mit einer Schwächung des anderen Maßes einher. Empfiehlt man zum Beispiel eine größere Anzahl an Objekten, steigt die Chance relevante Objekte zu empfehlen und der Recall steigt. Allerdings werden gleichzeitig sehr wahrscheinlich auch mehr Objekte empfohlen, die nicht relevant sind, d.h. die Precision wird zurückgehen. Ein Maß, das Precision und Recall in einem Wert vereint, ist die F1-Metrik [Herlocker et al., 2004].

$$F1 = 2 * \frac{P * R}{P + R} \tag{3.9}$$

Je nach Wichtigkeit der Precision oder des Recall kann in der F-Metrik auch eine andere Gewichtung als die Gleichgewichtung beim F1-Maß gewählt werden.

3.6.3 Weitere Metriken

Mit der fortschreitenden Entwicklung der Algorithmen zur Empfehlungsgenerierung wurde schnell klar, dass der Wert einer Empfehlung für seine Zielperson nicht nur von einer möglichst korrekt vorhergesagten Bewertung abhängt. Vor allem inhaltsbasierte Systeme haben den Nachteil, dass sie zwar sehr gut für bekannte Nutzerpräferenzen passende Empfehlungen treffen können. Dies führt allerdings häufig dazu, dass lediglich mehr vom Gleichen und damit häufig weniger Nützliches

präsentiert wird. Aus diesem Grund wurden mit der Zeit neue Evaluationskriterien definiert, die weitere Kriterien bzgl. der Nutzererfahrung berücksichtigen. Ein Beispiel sind Coverage und Serendipity [Herlocker et al., 2004, Ge et al., 2010, Kaminskas und Bridge, 2016].

Coverage Die Coverage eines Empfehlungssystems bzgl. empfohlener Objekte kann auf zwei Arten evaluiert werden. Die *Prediction Coverage* berechnet den Anteil I_p aller Objekte I, für die das System bei einer Empfehlungsgenerierung eine fundierte Aussage über ihre Relevanz treffen kann. Von einer „fundierten" Aussage kann man sprechen, wenn zum Beispiel bei einem bewertungsbasierten Ansatz genügend Bewertungen für ein Objekt vorliegen, um Ähnlichkeiten berechnen und Bewertungen abschätzen zu können. Bei einem regelbasiertem System kann ein Objekt dann gut eingeschätzt werden, wenn für es anwendbare Regeln vorhanden sind.

$$PredictionCoverage = \frac{|I_p|}{|I|} \tag{3.10}$$

Die *Catalog Coverage* berechnet den Anteil der Objekte, die in einem untersuchten Zeitraum tatsächlich empfohlen wurden. Diese Metrik ist vor allem für Empfehlungssysteme geeignet, in denen nicht eine einzelne Empfehlung, sondern eine Top-N Liste an Empfehlungen generiert wird.

$$CatalogCoverage = \frac{\left|\cup_{j=1\ldots N} I_L^j\right|}{|I|} \tag{3.11}$$

Es ist allerdings nicht sinnvoll, nur möglichst viele verschiedene Objekte zu empfehlen, ohne auf die tatsächliche Relevanz der Objekte für die Nutzer zu achten. Deswegen gibt es für beide Metriken gewichtete Varianten, die auch die Nützlichkeit der Empfehlungen berücksichtigen [Herlocker et al., 2004]. Die Definition von „Nützlichkeit" hängt wiederum von der Anwendung und dem Ziel der Evaluation ab. Nützliche Empfehlungen können zum Beispiel Empfehlungen mit hohen Bewertungen sein, aber auch Empfehlungen für neuartige Objekte, die bisher möglicherweise nur wenige Bewertungen erhalten haben.

Eine weitere Metrik ist die *User Coverage* [Jannach et al., 2010]. Sie berechnet den Anteil U_p aller Nutzer U, für die Empfehlungen ausgesprochen werden können. Sie ist vor allem für die Analyse der Systemperformance in Situationen mit neuen Nutzern interessant. Auch bei der User Coverage ist es ratsam, die Güte der Empfehlungen zu berücksichtigen, um aussagekräftigere Ergebnisse zu erhalten. Beim kollaborativen Filtern wird nämlich zum Beispiel die durchschnittliche Bewertung einer Person als Vorhersage zurückgegeben, falls aufgrund mangelnder Bewertungen keine Ähnlichkeit zu anderen Nutzern berechnet werden kann, siehe Gleichung (3.3). Dadurch können zwar mit einem zufriedenstellenden Fehler (z.B. MAE) Bewertungen vorhergesagt werden, die Qualität der Empfehlungen leidet allerdings.

$$UserCoverage = \frac{|U_p|}{|U|} \tag{3.12}$$

Serendipity Im Gegensatz zu anderen Evaluationsmetriken ist die *Serendipity* ein äußerst subjektives Maß. Herlocker und Kollegen [Herlocker et al., 2004] definierten sie als das Ausmaß, in welchem ein empfohlenes Objekt sowohl interessant als auch überraschend ist. Die Herausforderung besteht darin, eine Überspezifikation der Nutzermodelle und die Empfehlung „offensichtlicher" Objekte zu vermeiden und dennoch die Gefahr nicht zufriedenstellender und nutzloser Empfehlungen zu minimieren. Ge und Kollegen [Ge et al., 2010] definierten Serendipity folgendermaßen:

$$Serendipity = \frac{\sum_{i=1}^{N} u(RS_i)}{N} \tag{3.13}$$

RS_i steht dabei für ein Objekt, das als überraschend bzw. unerwartet bezeichnet werden kann. Um diese Objekte zu finden, verglichen Ge und Kollegen die Empfehlungen eines primitiven Modells, das eine große Vorhersehbarkeit hatte, mit den Empfehlungen des zu evaluierenden Empfehlungssystems. N ist die Anzahl der Objekte, die bei diesem Vergleich als „unerwartet" erkannt wurden. $u(RS_i)$ beschreibt die Nützlichkeit des jeweiligen Objekts (nützlich= 1, nutzlos= 0), die wiederum vom subjektiven Feedback der Nutzer wie zum Beispiel Bewertungen abhängt.

Alternativ zur Serendipity können auch die *Neuartigkeit* und *Vielfalt* von Empfehlungen untersucht werden. Die Neuartigkeit unterscheidet sich von der Serendipity dahingehend, dass in beiden Fällen von den Nutzern zwar überraschende Objekte entdeckt werden, dass bei der Neuartigkeit allerdings die Nutzer selbst die Objekte (zum Beispiel in einer Liste) entdecken und nicht durch das System darauf hingewiesen werden. Von einer erhöhten Vielfalt kann dagegen gesprochen werden, wenn neue Daten im Nutzermodell die Empfehlung neuer Objekte ermöglichen. Die Zielperson der Empfehlung könnte dann allerdings weniger von den neuen Empfehlungen überrascht sein.

4 Empfehlungsauswahl

Die Hauptaufgabe eines jeden Empfehlungssystems ist die situative Auswahl der nützlichsten Empfehlung(en) für eine Zielperson. Je besser diese Aufgabe bewältigt wird, desto wahrscheinlicher ist es, dass Nutzer Empfehlungen annehmen und Vertrauen in die Kompetenz des Systems aufbauen. Wie die Verhaltensmodelle in Kapitel 2.1 gezeigt haben, ist eine qualitativ hochwertige Empfehlungsauswahl für beratende Empfehlungssysteme aber auch deswegen wichtig, da Empfehlungen, die nicht zur aktuellen Lage der Nutzer (u.a. Bedürfnisse, Motivation, Fähigkeiten) passen, nur geringe Erfolgschancen haben. Das häufig in Empfehlungssystemen auftretende Cold-Start-Problem erschwert die Empfehlungsauswahl zusätzlich, siehe Kapitel 3.2.

Forschungsfragen Um in assistierenden Empfehlungssystemen auch in Phasen mit wenigen Nutzerbewertungen eine qualitativ hochwertigere Empfehlungsauswahl erreichen zu können, wird in dieser Dissertation die Nutzung zusätzlicher Nutzermodelle untersucht, mit denen der Mangel an Bewertungen ausgeglichen werden kann. Für die Beschreibung der systemrelevanten, aktuellen Eigenschaften der Nutzer erscheinen anwendungsspezifische und theoriebasierte Nutzermodelle aus den Sozialwissenschaften als vielversprechend. Hat man ein geeignetes Nutzermodell identifiziert, stellt sich jedoch die Frage, wie es sich in ein klassisches, bewertungsbasiertes Filterverfahren wie das kollaborative Filtern integrieren lässt. Außerdem ist unklar, ob durch die Integration eines theoriebasierten Nutzermodells die Qualität der Empfehlungen tatsächlich verbessert werden kann.

Zunächst werden in Kapitel 4.1 Arbeiten vorgestellt, die bereits erfolgreich psychologische Theorien für die Empfehlungsauswahl eingesetzt haben. Anschließend werden vielversprechende Modelle für den Einsatz in CARE und SavER beschrieben, siehe Kapitel 4.2. Diese Modelle wurden auf verschiedene Art in einen klassischen kollaborativen Filteralgorithmus integriert und die Qualität der neuen Algorithmen evaluiert, siehe Kapitel 4.3. Zusätzlich zur Beantwortungen der Forschungsfragen, werden in Kapitel 4.4 der Ablauf und die Erkenntnisse der nutzerzentrierten Entwicklung eines prototypischen CARE-Systems beschrieben. Dadurch soll beispielhaft gezeigt werden, welche Anforderungen mögliche Nutzer eines beratenden Empfehlungssystems an die Empfehlungsauswahl haben können und wie diese bei der Umsetzung des Systems berücksichtigt werden können. Kapitel 4.5 fasst die Forschungsarbeit bzgl. der Empfehlungsauswahl zusammen.

4.1 Theoriebasierte Empfehlungsauswahl

Um trotz der bekannten Schwächen (z.B. Cold-Start-Problem) traditioneller Filtertechniken qualitativ hochwertige Empfehlungen erreichen zu können, wurde schon häufiger der Einfluss psychologischer und sozialwissenschaftlicher Faktoren auf die menschliche Entscheidungsfindung untersucht, siehe [Nunes, 2010] für eine ausführ-

liche Diskussion entsprechender Arbeiten. Die im Folgenden vorgestellten Arbeiten wurden unter dem Gesichtspunkt ausgewählt, dass sie erfolgreich validierte soziologische und psychologische Modelle für die Empfehlungsauswahl genutzt haben.

Demographie Bereits 1979 untersuchte Rich [Rich, 1979] Stereotypen, um Literaturempfehlungen personalisieren zu können. Die Annahme war, dass Personen, die anhand ihrer Demographie dem gleichen Stereotypen zugeordnet werden können, auch ähnliche Vorlieben für bestimmte Objekte haben. Typische Informationen zur Unterscheidung der Nutzer sind u.a. Geschlecht, Alter, Herkunft, Bildungsstand oder Einkommen. Diese Daten können durch explizite Dialoge mit den Nutzern [Krulwich, 1997] oder durch die Analyse veröffentlichter persönlicher Daten [Pazzani, 1999] gewonnen werden. Gute Beispiele für die Kombination demographischer Daten und kollaborativer Filtertechniken sind die Arbeiten von Pazzani [Pazzani, 1999] und Koren und Kollegen [Koren et al., 2009].

Persönlichkeit Erkenntnisse aus der Psychologie haben gezeigt, dass die Persönlichkeit beständig und als einer der wichtigsten Faktoren überhaupt das menschliche Verhalten beeinflusst und eine starke Verbindung zwischen der Persönlichkeit und Vorlieben und Interessen besteht [Jung und Baynes, 1923]. Durch diese Erkenntnisse bestärkt wurde auch bereits erforscht, wie die Persönlichkeit der Nutzer für die Empfehlungsauswahl genutzt werden könnte [Dunn et al., 2009, Hu und Pu, 2011, Karumur et al., 2016, Nunes, 2008].

Hu und Pu [Hu und Pu, 2011] versuchten durch die Berücksichtigung der Persönlichkeit der Nutzer das Cold-Start-Problem eines Musik-Empfehlungssystems zu mildern. Hierfür verwendeten sie das *Big-Five-Persönlichkeitsmodell*, auf das in Kapitel 5.3.1 im Rahmen der Generierung personalisierter Empfehlungstexte näher eingegangen wird. Für die Integration der Persönlichkeit in ein kollaboratives Empfehlungssystem entwickelten Hu und Pu drei Verfahren: Ein rein auf die Persönlichkeit basierender Ansatz, ein Hybrid, der die bewertungsbasierte Ähnlichkeit der Nutzer und ihre Ähnlichkeit basierend auf der Persönlichkeit linear kombinierte und eine weitere hybride Kombination in Form einer Kaskade. Alle drei Verfahren erreichten in Cold-Start-Szenarien signifikant bessere Ergebnisse als ein klassischer kollaborativer Filter. Speziell der kaskadierende Ansatz führte zu einer stark verbesserten Qualität, sowohl im Hinblick auf die Genauigkeit der vorhergesagten Bewertungen als auch hinsichtlich der Klassifikation in „relevant“ und „nicht relevant“.

Emotionen Information über den emotionalen Zustand von Nutzern wurden ebenfalls bereits zur Auswahl von Empfehlungen genutzt. Aufgrund von Sensordaten wie Herzrate, Temperatur oder Hautleitwert schlossen Nasoz und Kollegen [Nasoz et al., 2010] auf die aktuelle Emotion von Autofahrern. Sie fokussierten sich auf die negativen Emotionen Wut, Frustration, Panik, Langeweile und Schläfrigkeit, da diese zu gefährlichen Situationen im Straßenverkehr führen können. Außerdem wurden Faktoren wie das Alter, das Geschlecht und die Persönlichkeit der Fahrer, die

das Fahrverhalten ebenfalls beeinflussen können, berücksichtigt. In sicherheitskritischen Situationen sollte das System versuchen, den emotionalen Zustand der Fahrer zu stabilisieren und so das Sicherheitsrisiko zu reduzieren. Beispiele für Maßnahmen waren das Wechseln des Radiosenders, das Öffnen des Fensters oder Empfehlungen, eine Pause einzulegen oder Entspannungsübungen durchzuführen.

Auch ein 2016 gestartetes Projekt namens „I-CARE“ berücksichtigt die Emotionen der Nutzer bei der Entscheidungsfindung. Das Ziel des Projekts ist die Aktivierung an Demenz erkrankter Menschen. Durch eine Analyse der Mimik, der Stimme und der Bewegungen der Patienten soll auf ihre Emotionen geschlossen werden. Anschließend soll durch eine gezielte Auswahl von Aktivierungsinhalten besser auf die „Tagesform“ der Patienten eingegangen werden können.[11]

Vertrauen Durch die Berücksichtigung der Vertrauenswürdigkeit von Nutzern kann die Genauigkeit der Empfehlungsauswahl ebenfalls verbessert werden. Im Grunde geht es darum, die Vertrauenswürdigkeit der Nutzer bzgl. der Bewertung von Objekten zu beurteilen und anschließend nur vertrauenswürdige Nutzer für die kollaborative Auswahl von Empfehlungen zu berücksichtigen. Die Beurteilung der Vertrauenswürdigkeit kann in zweierlei Hinsicht geschehen.

Die erste Möglichkeit besteht darin, die abgegebenen Bewertungen einer Person mit den abgegebenen Bewertungen der anderen Nutzer zu verglichen [O'Donovan und Smyth, 2005]. Je höher die Übereinstimmung ausfällt, umso vertrauenswürdiger bzw. kompetenter ist die Person. Beruht die Einschätzung der Nutzer jedoch nur auf unzureichenden Daten (z.B. in Cold-Start-Szenarien), sind die Ergebnisse nur wenig zuverlässig.

Alternativ können von den Nutzern explizite Bewertungen für die Vertrauenswürdigkeit anderer Personen abgefragt werden [Massa und Avesani, 2007a, Rafailidis und Crestani, 2017]. So entsteht ein Netzwerk mit individuellen Vertrauensbeziehungen zwischen Nutzern. Durch Vertrauensmetriken ist es anschließend sogar möglich, auch für zwei Nutzer, für die bisher keine direkte Vertrauensverbindung besteht, eine Aussage über ihre gegenseitige Vertrauenswürdigkeit zu treffen. Globale Metriken berechnen jedoch nur eine durchschnittliche Bewertung der Vertrauenswürdigkeit (Reputation) einer Person basierend auf allen abgegebenen Vertrauensbewertungen bzgl. dieser Person [Page et al., 1998]. Lokale Metriken dagegen treffen personalisierte Annahmen über das Vertrauensverhältnis zwischen Personen [Golbeck, 2005, Massa und Avesani, 2007b]. Hierfür machen sie sich die Transitivität von Vertrauen (A vertraut B und B vertraut C. Dann kann auch A C vertrauen.) zu Nutzen.

Sowohl für die implizite [O'Donovan und Smyth, 2005] als auch für die explizite Variante [Massa und Avesani, 2007a], die Vertrauenswürdigkeit von Personen einzuschätzen, zeigte sich, dass eine Berücksichtigung der ermittelten Vertrauenswerte die Empfehlungsauswahl signifikant verbessern kann. Massa und Avesani [Massa und Avesani, 2007a] erzielten in Cold-Start-Szenarien zum Beispiel durch die

[11] https://www.projekt-i-care.de/

Verwendung einer lokalen Vertrauensmetrik, die nur Nutzer mit direkten Vertrauensverbindungen berücksichtigte, die besten Ergebnisse. Die Einbeziehung weiterer Nutzer, zu denen im Vertrauensnetzwerk nur indirekt eine Verbindung zur Zielperson bestand, konnte zwar die Prediction Coverage verbessern, führte jedoch mit steigender Schrittweite im Netzwerk (maximal erlaubte Entfernung zwischen zwei Personen) zu einem wachsenden Fehler bei der Vorhersage der Bewertungen. Allerdings übertrafen selbst Verfahren mit größeren Schrittweiten bei schwer vergleichbaren Nutzern die Genauigkeit des traditionellen kollaborativen Filterns.

Die Ergebnisse einer eigenen Studie, in der die UX von Reputationssystemen mit verschiedenen Vertrauensmetriken verglichen wurde, können in [Hammer et al., 2013] nachgelesen werden.

Die eben vorgestellten Arbeiten zeigen, dass die Qualität traditioneller Filtertechniken durch die Einbeziehung geeigneter Theorien und Modelle gesteigert werden kann. Aus diesem Grund war die Annahme der in diesem Kapitel beschriebenen Forschungsarbeit, dass auch die Empfehlungsauswahl in beratenden Empfehlungssystemen durch geeignete theoriebasierte Nutzermodelle verbessert werden kann. Geeignete Nutzermodelle für CARE und SavER, ihre Integration in die traditionelle kollaborative Filtertechnik sowie die Evaluation der entstandenen Filterverfahren werden in den folgenden Kapiteln detailliert beschrieben.

4.2 Theoriebasierte Nutzermodelle in CARE und SavER

4.2.1 CARE - Wohlbefinden

Da Bewertungen früherer Empfehlungen durch Schwankungen des Wohlbefindens von Senioren nicht in jeder Situation gleich aussagekräftig sind, benötigt das CARE-System für eine qualitativ hochwertige Empfehlungsauswahl zusätzlich Wissen über das aktuelle Wohlbefinden der Zielperson sowie den aktuellen Umgebungskontext.

Das von der New Economics Foundation (nef) erstellte „National Accounts of Well-being Framework“ [Michaelson et al., 2009] unterscheidet verschiedene Faktoren und Unterfaktoren von Wohlbefinden, siehe Abbildung 4.1. Die Faktoren in der untersten Schicht dieser Hierarchie können auf konkrete Fragen in einem Fragebogen übertragen werden, um eine subjektive Messung des Wohlbefindens zu ermöglichen.

Ein weiterer Ansatz zur Unterteilung von Wohlbefinden ist der „Index of Well-Being in Older Populations“ des Stanford Center on Longevity (SCL) und des Population Reference Bureau (PRB). Dieser Index unterscheidet *materielles*, *physikalisches* und *kognitives Wohlbefinden* sowie *soziales Engagement* [Kaneda et al., 2011].

Im CARE-Projekt wurde durch die Anpassung bestehender Modelle ein Modell entwickelt, das nur Kategorien und Subkategorien enthält, für die Empfehlungen ausgesprochen werden können. Zur Identifikation dieser Kategorien, wurden strukturierte Interviews mit 21 Personen (13 Frauen, acht Männer) im Alter von 67 bis 83 Jahren durchgeführt [Rist et al., 2015]. Dabei wurde der Fragebogen des „National Accounts of Well-being Framework“ verwendet, um die Lebenssituation, Probleme

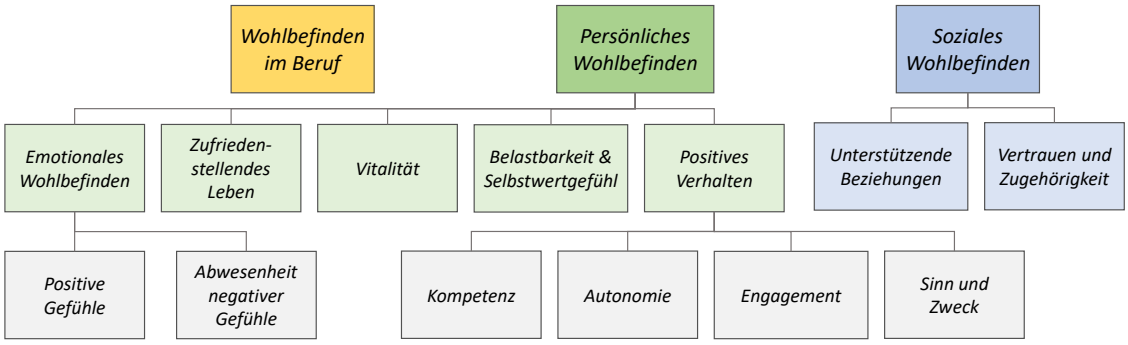

Abbildung 4.1: Struktur und Faktoren des National Accounts of Well-being Framework (nach [Michaelson et al., 2009])

und Bedürfnisse der Senioren abzufragen. Es zeigte sich, dass vor allem die körperliche Vitalität, beeinflusst durch sportliche Aktivitäten und eine gesunde Ernährung, die mentale Leistungsfähigkeit, das emotionale Wohlbefinden sowie soziale Kontakte eine wichtige Rolle im Leben der Teilnehmer spielten. Außerdem stellte sich heraus, dass auch die Umgebung, d.h. die Gegebenheiten zuhause und am Wohnort, einen starken Einfluss auf das subjektive Wohlbefinden hat. Diese Faktoren werden im Folgenden detailliert erläutert. Das entstandene Modell, inklusive der Subkategorien, ist in Abbildung 4.2 zu sehen.

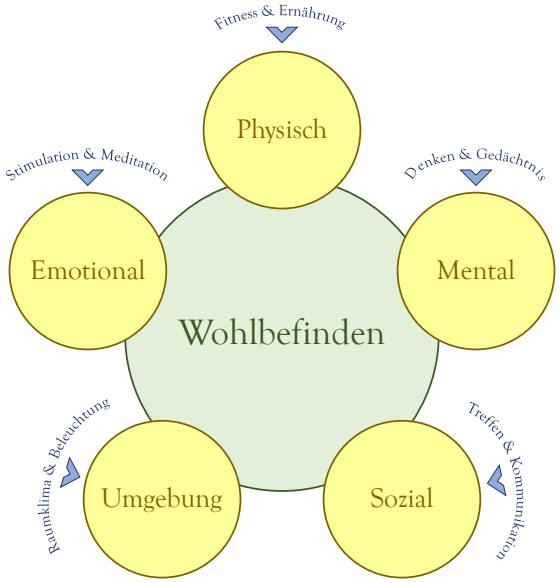

Abbildung 4.2: Modell für „Wohlbefinden“ aus dem Projekt CARE [Rist et al., 2015]

Physisches Wohlbefinden Häufig mangelt es älteren Menschen aufgrund von Krankheiten oder altersbedingten Gebrechen an körperlicher Fitness. Durch körperliche Übungen und eine gesunde Ernährung kann die Physis der Senioren und damit auch ihre Selbstständigkeit und Mobilität verbessert werden.

Mentales Wohlbefinden Zu den mentalen Fähigkeiten zählen u.a. das Erinnerungsvermögen, kreatives Handeln, logisches Denken und das Ziehen von Schlüssen. All diese Fähigkeiten können mit zunehmendem Alter abnehmen. Sie können durch geeignete Übungen wie zum Beispiel Gedächtnisübungen und Denksportaufgaben allerdings auch trainiert werden.

Soziales Wohlbefinden Soziale Isolation ist ein weiteres Problem vieler Senioren. Bei Defiziten in dieser Kategorie sollte versucht werden, regelmäßige Treffen oder Telefonate mit Freunden, Verwandten oder Bekannten anzustoßen. Auch eine Verbesserung der Mobilität und die Bekämpfung von Müdigkeit und Lustlosigkeit im Rahmen anderer Kategorien des Wohlbefindens können das soziale Leben älterer Menschen befeuern.

Emotionales Wohlbefinden Ausgelöst durch die bereits genannten Probleme können sich Senioren niedergeschlagen und deprimiert fühlen. Außerdem könnte emotionaler Stress andere wichtige Faktoren wie zum Beispiel den Schlaf beeinträchtigen. Das Ziel sollte u.a. sein, die Frequenz und Intensität positiver Gefühle zu steigern und negative Gefühle zu vermeiden. Dies kann u.a. durch mediale Stimulationen und Entspannungsübungen erreicht werden.

Umgebung Befinden sich das Zuhause und die Umgebung einer älteren Person in einem guten Zustand, so wirkt sich dies auch positiv auf das allgemeine Wohlbefinden der Person aus. Ein gutes Raumklima kann zum Beispiel vitalisierend wirken. Angenehme Lichtverhältnisse oder eine schön gestaltete Einrichtung können dagegen eine ausgleichende Wirkung haben.

Erfassung von Wohlbefinden Zur subjektiven Einschätzung des Wohlbefindens gibt es eine große Anzahl an unterschiedlichen Fragebögen. Neben dem „National Accounts of Well-being Framework“ [Michaelson et al., 2009] sind auch die Fragebögen der World Health Organization (WHO) für unterschiedliche Zielgruppen und Anforderungen sehr interessant. Besonders relevant erscheint zunächst der WHOQOL-AGE [Caballero et al., 2013], der sich direkt auf das Wohlbefinden älterer Menschen bezieht. Allerdings deckt dieser Fragebogen nicht alle Kategorien des Modells in CARE ab. Der WHOQOL-Bref [Skevington et al., 2004] passt dagegen mit seiner Unterteilung in die Domänen „Allgemeine Lebensqualität“, „Körperliche Gesundheit“, „Psychologisches Wohlbefinden“, „soziale Beziehungen“ und „Umfeld/Umgebung“ sehr gut zu der Modellierung des Wohlbefindens in CARE. Eine weitere Alternative wäre der WHOQOL-100 [Power et al., 1999]. Allerdings ist dieser Fragebogen mit seinen 100 Fragen im Vergleich zum WHOQOL-Bref (26 Fragen) nicht geeignet, um eine regelmäßige Erfassung des Wohlbefindens der Nutzer durchzuführen. Ein Vorteil aller WHOQOL-Fragebögen ist, dass ihre Auswertung numerische Werte für das generelle Wohlbefinden und die abgefragten Kategorien

liefert. Zusammengefasst in einem Nutzermodell lassen sich diese Werte gut in die Berechnungen während der Empfehlungsauswahl integrieren.

Empfehlungen für Wohlbefinden Während des CARE-Projekts wurde eine Datenbank mit einer Vielzahl von Empfehlungen für alle Kategorien des Wohlbefindens erstellt. Dazu gehörten u.a. Aktivitäten, die ein Teil der Teilnehmer der Interview bereits regelmäßig ausübte und als wichtig bezeichnete. Weitere Quellen waren verschiedene Ratgeberseiten, wie die der Apotheken Umschau[12] und des Senioren Ratgebers [13]. Abbildung 4.3 zeigt Beispiele für Empfehlungen aus den verschiedenen Kategorien. Zur besseren Unterscheidung der Aktivitäten im Empfehlungssystem wurden die beiden Subkategorien *Übungen* und *Ernährung* des *physischen Wohlbefindens* wie getrennte Kategorien behandelt.

Abbildung 4.3: Beispielhafte Empfehlungen der Kategorien emotionales Wohlbefinden, physisches Wohlbefinden (Übungen), physisches Wohlbefinden (Ernährung), soziales Wohlbefinden, Umgebung und mentales Wohlbefinden

Insgesamt entstanden im Laufe des Projektes 126 Empfehlungen, die alle Kategorien des Wohlbefindens abdeckten. Der größte Anteil an Empfehlungen (35) vereinte sich in der Kategorie für mentales Wohlbefinden. Hierzu zählten Rätselaufgaben und Wissensfragen aus dem privaten und Allgemeinwissensbereich. Im Gegensatz zu anderen Empfehlungen waren die Aufgaben dieser Kategorie interaktiv gestaltet. Die Senioren konnten entweder eine von mehreren Antworten auswählen oder Antworten wie zum Beispiel erfragte Telefonnummer direkt eingeben, siehe Abbildung 4.3 (rechts unten). Außerdem enthielt die Datenbank noch viele Empfehlungen (33) für Übungen zur Lockerung und Stärkung aller Körperregionen. Diese Empfehlungen waren teilweise animiert, um den Ablauf einer Übung besser darstellen zu können. Eine weitere große Menge an Empfehlungen (27) gehörte zur Kategorie des emotionalen Wohlbefindens. Es gab Empfehlungen für Entspannungs- und Meditationsübungen sowie Scherzfragen zur Aufheiterung. Wie in Abbildung 4.3 (links oben)

[12] http://www.apotheken-umschau.de/ [13] http://www.senioren-ratgeber.de

zu sehen ist, waren die Meditationsübungen etwas atmosphärischer gestaltet, um bereits durch die Gestaltung der Empfehlung eine gewisse beruhigende Wirkung zu erzielen. Empfehlungen hinsichtlich der Umgebung beschränkten sich auf die Pflege des eigenen Umfelds (z.B. Lüften, Gießen) und Empfehlungen zur Beobachtung der Natur. Aus diesem Grund waren für diese Kategorie lediglich zwölf verschiedene Aktivitäten vorhanden. Ähnlich verhielt es sich für die Ernährungstipps des Systems (11). Diese zielten nicht auf konkrete Rezepte ab, sondern auf grundlegende Regeln wie die Aufnahme von Vitaminen oder regelmäßiges Trinken. Für das soziale Wohlbefinden wurden acht allgemeine Empfehlungen für gemeinsame Aktivitäten wie Kochen, Spazieren gehen oder alte Bilder betrachten erstellt.

4.2.2 SavER - Energieverhalten

Stephenson und Kollegen [Stephenson et al., 2010] versuchten in ihrem konzeptuellen Modell der *Energy Cultures* zu modellieren, wie unterschiedliches Energieverhalten von Menschen entsteht und wie es von Außen beeinflusst werden kann. Sie identifizierten drei Faktoren, die das Energieverhalten von Menschen beeinflussen, siehe Abbildung 4.4.

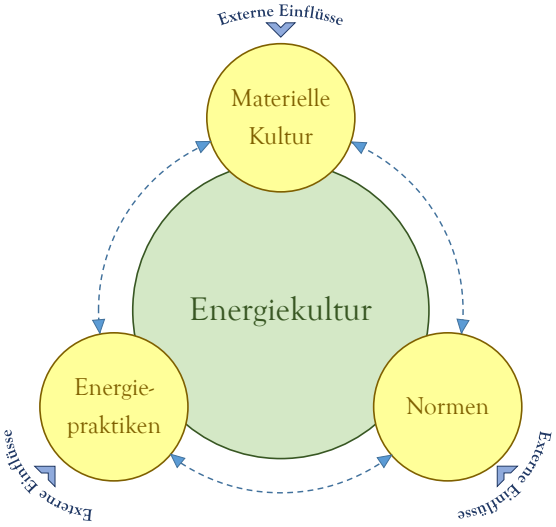

Abbildung 4.4: Energy Cultures Framework (nach [Stephenson et al., 2010])

Materielle Kultur Die *materielle Kultur* umfasst alle physikalischen Objekte und Gegebenheiten, die das Energieverhalten einer Person beeinflussen. Für die Beurteilung der materiellen Kultur werden u.a. die Art und der Zustand der Isolierung, der Stromquellen und der Heizung der Wohnung bzw. des Hauses berücksichtigt. Außerdem sind der Besitz elektrischer Geräte wie z.B. Fernseher, Trockner oder Waschmaschine sowie vorhandene Fortbewegungsmittel wie Autos, motorisierte Zweiräder oder Fahrräder ausschlaggebend.

Normen Die *Normen* einer Person setzen sich aus Meinungen und Einstellungen der Person und ihres näheren Umfeldes zu bestimmten Fragestellungen im Bereich Energieverbrauch zusammen: Sind Umweltschutz und Energiesparen wichtige Themen? Welche Zimmertemperatur gilt als angemessen? Sollte man sich lieber wärmer anziehen oder die Heizung hochdrehen? Hat beim Einkauf von Geräten deren Energieeffizienzklasse einen Einfluss auf die Kaufentscheidung? Welche Fortbewegungsmittel sollten genutzt werden?

Energiepraktiken Die *Energiepraktiken* einer Person beinhalten Tätigkeiten, die den Energieverbrauch beeinflussen. Dazu gehören zum einen einmalige Aktionen wie die Anschaffung von Geräten oder bauliche Veränderungen und zum anderen regelmäßige Aktivitäten wie die Nutzung öffentlicher Verkehrsmittel oder das Ausschalten des Stand-By Modus elektrischer Geräte.

Die Faktoren beeinflussen und verstärken sich gegenseitig. Eine Person, für die Energiesparen nicht wichtig ist, wird auch weniger auf ihren Energieverbrauch achten. Ein gut isoliertes Zuhause kann andererseits dazu führen, dass die Bewohner allgemein weniger heizen. Die dadurch entstandenen finanzielle Einsparungen könnten dann wiederum zu einem Umdenken beim Thema Energiesparen führen.

Auch äußere Einflüsse können zu einer Änderung des Energieverhaltens führen. Ein besonders drastisches Beispiel hierfür ist das Unglück in Fukushima[14], dass bei vielen Menschen zu einem Umdenken geführt hat. Das Energieverhalten kann aber auch gezielt durch Gesetze, finanzielle Förderungen oder erhöhte Strompreise gesteuert werden. Allerdings ist es schwer und oft wenig effektiv, ein einheitliches Vorgehen bzw. einheitliche Richtlinien zur Förderung von Verhaltensänderungen zu definieren. Sowohl das Energieverhalten als auch die Voraussetzungen für ein energiesparendes Verhalten sind dafür innerhalb der Population zu heterogen.

Geht man jedoch davon aus, dass Menschen mit einer ähnlichen Energiekultur ähnliche Interessen bzgl. des Themas Energiesparen haben, wäre es möglich, auf die jeweiligen Gruppen gezielt einzugehen. Bisher negativ beurteilte Faktoren bieten dabei ein größeres Potential zur Steigerung der Energieeffizienz als bereits positiv bewertete Faktoren. Eine mögliche Kategorisierung von Energieverhalten gelang Lawson und Williams [Lawson und Williams, 2012]. Sie identifizierten mittels einer Umfrage mit über 2.000 Teilnehmern vier verschiedene Energiekulturen.

Energy Economic Personen dieser Gruppe sind jung und haben nur geringe finanzielle Mittel. Sie leben in kleinen, bescheidenen Mietwohnungen, die sich häufig in schlecht isolierten Häusern mit veralteten Heizsystemen befinden. Aus diesem Grund spielen Sparsamkeit und energiesparendes Verhalten eine wichtige Rolle. Energieökonomische Personen sind vor allem an Praktiken interessiert, mit denen sie ohne größeren finanziellen Aufwand Energie einsparen und langfristig ihre materiellen Verhältnisse verbessern können.

[14] http://www.spiegel.de/thema/fukushima/

Energy Extravagant Diese Gruppe umfasst Menschen, die ihre Lebensqualität über die Energieeffizienz stellen. Sie stammen meist aus Familienhaushalten mit hohem Einkommen. Die finanziellen Mittel werden u.a. in große Häuser und moderne Technologien investiert. Energieeinsparungen durch diese Technologien (z.B. moderne Heizsysteme) sind oft aber nur ein positiver Nebeneffekt. Eine Änderung des Energieverhaltens würde zunächst voraussetzen, dass das Bewusstsein und Interesse für energiesparendes Verhalten geweckt wird.

Energy Efficient Teil dieser Energiekultur sind vor allem ältere Personen, die sich teilweise bereits im Ruhestand befinden und deren Kinder bereits ausgezogen sind. Sie schätzen vor allem zweckmäßige Dinge. Deswegen besitzen sie wenige elektronische Geräte oder benutzen diese selten. Außerdem leben sie in gut isolierten Wohnungen oder Häusern und zeichnen sich auch sonst durch ein energiesparendes Verhalten aus. Von besonderem Interesse sind für diese Gruppe Energiesparpraktiken, die keine zusätzlichen Kosten verursachen.

Energy Easy Mitglieder dieser Energiekultur sind mittleren bis fortgeschrittenen Alters und haben hohe Einkommen. Sie bevorzugen einen möglichst einfachen und angenehmen Lebensstil und zeigen dementsprechend keinerlei Interesse Energie einzusparen. Zum Beispiel heizen sie immer ihr komplettes Haus. Auch im Falle dieser Energiekultur wäre zunächst ein Umdenken von Nöten, ehe aufwendigere Energiesparpraktiken in Frage kommen.

4.3 Theoriebasierte Nutzermodelle in kollaborativen Filtern

Die folgenden Varianten des kollaborativen Filterns orientieren sich an den in Kapitel 4.1 vorgestellten Arbeiten. Sie ersetzen oder ergänzen den bewertungsbasierten Ansatz zur Ähnlichkeitsberechnung durch eine Einschätzung der Ähnlichkeiten, die auf anwendungsspezifischen theoriebasierten Nutzermodellen beruht.

Theoriebasiertes Nutzermodell - CARE Im CARE-Szenario wurde, anders als im gleichnamigen Projekt, eine leicht abgewandelte Variante des Modells aus Abbildung 4.2 eingesetzt. Die Kategorien dieses Modells bestehen aus den Domänen, die im WHOQOL-BREF [Skevington et al., 2004] abgefragt und bewertet werden. Der Vektor des Nutzermodells ist dementsprechend folgendermaßen aufgebaut:

$$\begin{pmatrix} LebensqualitaetAllgemein \\ KoerperlicheGesundheit \\ PsychologischesWohlbefinden \\ SozialeBeziehungen \\ BewertungdesUmfelds \end{pmatrix}$$

Der Vorteil dieser Anpassung ist, dass durch die Auswertung des WHOQOL-BREF-Fragebogens für jede Variable des Vektors direkt ein Wert zwischen 0 und 100 festgelegt werden kann [Skevington et al., 2004].

Theoriebasiertes Nutzermodell - SavER Das Nutzermodell im SavER-System leitet sich direkt vom Konzept der Energiekulturen ab.

$$\begin{pmatrix} Normen \\ MaterielleKultur \\ Energiepraktiken \end{pmatrix}$$

Aus Mangel an einem ausreichend kurzen Fragebogen zur Einschätzung der Energiekultur wurde in dieser Dissertation ein eigener Fragebogen erstellt. Basierend auf der Auswertung dieses Fragebogens erhalten alle Nutzer für jede Kategorie des Nutzermodells eine Bewertung zwischen 1 = „äußerst energieverschwenderisch“ und 5 = „äußerst energiesparend“.

Alternative kollaborative Filteralgorithmen Neben der rein theoriebasierten Ähnlichkeitsberechnung, die lediglich auf der Ähnlichkeit der anwendungsspezifischen Nutzermodelle beruht, wurden zum einen eine lineare Kombination der bewertungsbasierten und der theoriebasierten Ähnlichkeitsmaße und zum anderen eine hybride Variante mit Merkmalserweiterung untersucht. Bei der Merkmalserweiterung wird der theoriebasierte Ansatz zur Generierung künstlicher Bewertungen für eine anschließende bewertungsbasierte Filterung genutzt.

Im Folgenden wird der grobe Ablauf der verschiedenen Verfahren jeweils auch graphisch dargestellt. Abbildung 4.5 zeigt zum Beispiel den Ablauf einer klassischen kollaborativen Filterung, wie sie in Kapitel 3.2 erklärt wurde. Dunkle Kästchen kennzeichnen die Module, in denen Berechnungen durchgeführt werden. Helle Kästchen enthalten die jeweiligen Ein- und Ausgaben. Die durchnummerierten Übergänge beschreiben den jeweiligen Ablauf der verschiedenen Methoden zur Vorhersage der Bewertungen. Das bewertungsbasierte Nutzermodell wird in allen Varianten der Filterung für die Vorhersage der Bewertungen benötigt.

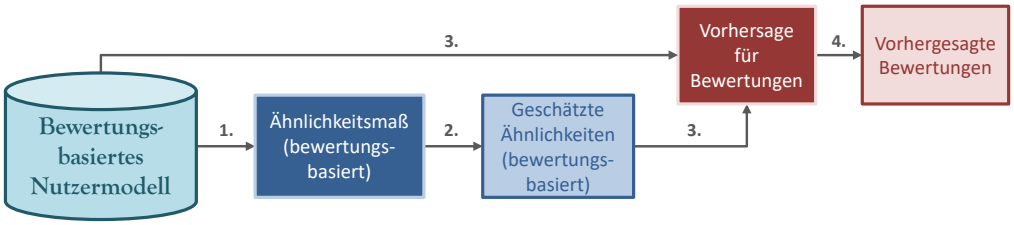

Abbildung 4.5: Ablauf der traditionellen kollaborativen Filterung

Theoriebasierte kollaborative Filterung Bei der rein theoriebasierten kollaborativen Filterung wird für die Bestimmung der Ähnlichkeit der Nutzer lediglich die Ähnlichkeit ihrer theoriebasierten Nutzermodelle berücksichtigt, siehe Abbildung 4.6. Hierfür wird wiederum die Pearson-Korrelation verwendet.

$$simM(u,v) = \frac{\Sigma_k(w_k(u) - \bar{w}(u))(w_k(v) - \bar{w}(v))}{\sqrt{\Sigma_k(w_k(u) - \bar{w}(u))^2 \Sigma_k(w_k(v) - \bar{w}(v))^2}} \tag{4.1}$$

$\bar{w}(u)$ und $\bar{w}(v)$ stehen für die durchschnittliche Bewertung, die die Nutzer u und v für die Dimensionen ihrer Nutzermodelle erreicht haben. $w_k(u)$ und $w_k(v)$ stehen für die Einschätzungen in den einzelnen Dimensionen.

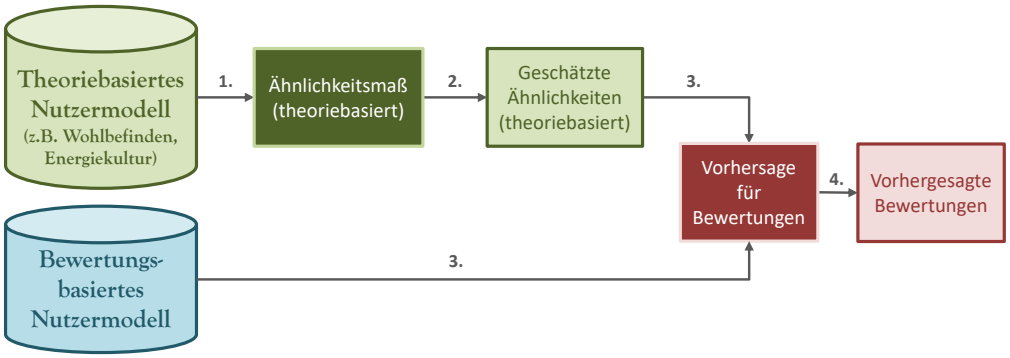

Abbildung 4.6: Ablauf der theoriebasierten kollaborativen Filterung

Linear hybride Filterung Die erste hybride Variante der Empfehlungsauswahl kombiniert die bewertungsbasierte und die theoriebasierte Ähnlichkeit auf lineare Weise, siehe Abbildung 4.7. Hierfür werden die bewertungsbasierte Korrelation aus Gleichung (3.1) und die theoriebasierte Korrelation aus Gleichung (4.1) durch eine lineare Gleichung zu einem neuen Ähnlichkeitsmaß kombiniert.

$$simL(u,v) = \alpha * sim(u,v) + (1-\alpha) * simM(u,v) \tag{4.2}$$

Mit der Variablen α kann gesteuert werden, wie stark die beiden Ähnlichkeitsmaße in die Berechnung des linear hybriden Ähnlichkeitsmaßes einfließen. Dieses α kann auch dynamisch (z.B. abhängig von der Menge der vorhandenen Bewertungen) gewählt werden [Hu und Pu, 2011]. Dadurch hat diese lineare Kombination im Vergleich zur Berechnung eines arithmetischen Mittelwerts oder eines harmonischen Mittelwerts, wie ihn O'Donovan und Smyth [O'Donovan und Smyth, 2005] verwendeten, den Vorteil, dass das jeweils zuverlässigere Ähnlichkeitsmaß stärker gewichtet werden kann.

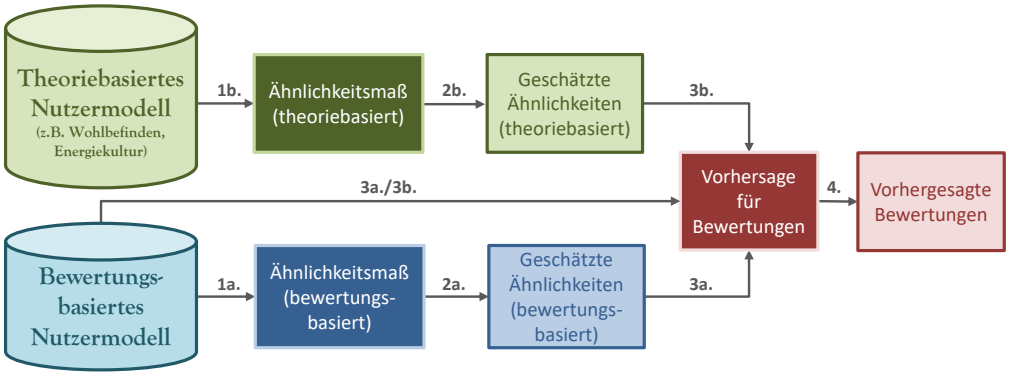

Abbildung 4.7: Ablauf der linear hybriden kollaborativen Filterung

Hybrides Filtern mit Merkmalserweiterung Eine weitere Variante die beiden Nutzermodelle zu kombinieren ist ein Ansatz mit Merkmalserweiterung, siehe Abbildung 4.8. Bei diesem Ansatz werden künstliche Bewertungen $\tilde{r}_{u,i}$ generiert, mit denen anschließend ein klassisches kollaboratives Filtern durchgeführt wird.

Die Erzeugung dieser künstlichen Bewertungen erfolgt durch das theoriebasierte kollaborative Filtern. Die Anzahl der ähnlichsten Nutzer, die für die Erzeugung künstlicher Bewertungen mit Gleichung (3.3) berücksichtigt werden, ist durch die frei wählbare Konstante β festgelegt. Der Vektor mit den Bewertungen des Nutzers u, der für die Pearson-Korrelation nach Gleichung (3.1) (Abbildung 4.8 - Schritt 6) verwendet wird, setzt sich dann folgendermaßen zusammen:

$$r'(u,i) = \begin{cases} r(u,i) & \text{falls Bewertung vorhanden} \\ \tilde{r}(u,i) \text{ mit Gleichung (3.3) und } simM & \text{sonst} \end{cases} \tag{4.3}$$

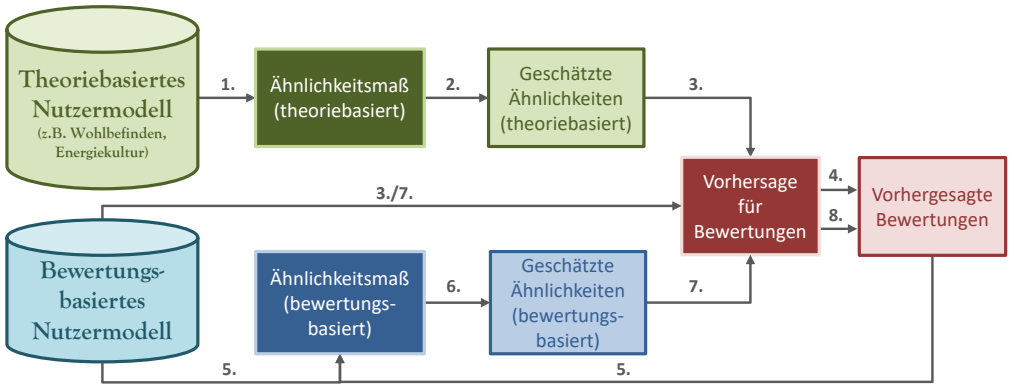

Abbildung 4.8: Ablauf der hybriden Filterung mit Merkmalserweiterung

4.3.1 Design der Evaluationen

Generierung von Datensätzen Da es weder für das CARE-System noch für das SavER-System Datensätze aus realen Systemen gibt, wurden in beiden Fällen zunächst in Umfragen Daten gesammelt. Die Datensammlungen enthielten zu allen Teilnehmern ein theoriebasiertes Nutzermodell sowie ein bewertungsbasiertes Nutzermodell mit Bewertungen für alle in der Umfrage präsentieren Empfehlungen.

Für die Evaluationen wurden die Datensätze in mehreren Durchläufen so in Trainings- und Testdatensätze zerlegt, dass Cold-Start-Problem-Szenarien mit wenigen Bewertungen für einen (New-User-Problem) oder alle Nutzer (Sparsity-Problem) untersucht werden konnten.

Hypothesen Sowohl für den Sparsity- als auch den New-User-Test wurde angenommen, dass die neuen Filterverfahren den traditionellen, nur auf Bewertungen basierenden kollaborativen Ansatz hinsichtlich der Qualität der Empfehlungen übertreffen. Als Qualitätskriterien wurden die Fähigkeit zur korrekten Vorhersage von

Bewertungen und die Fähigkeit zur Unterscheidung relevanter und irrelevanter Objekte untersucht. Zur Bestimmung der Genauigkeit der vorhergesagten Bewertungen wurde der Mean Absolute Error (MAE) berechnet. Für die korrekte Klassifikation der Objekte wurde das F1-Maß betrachtet. Die evaluierten Hypothesen lauteten:

1. Die Berücksichtigung theoriebasierter Nutzermodelle führt in Sparsity-Szenarien zu einer besseren Vorhersage zukünftiger Bewertungen.

2. Die Berücksichtigung theoriebasierter Nutzermodelle führt in Sparsity-Szenarien zu einer besseren Erkennung relevanter Aktionen.

3. In New-User-Szenarien führt die Berücksichtigung theoriebasierter Nutzermodelle zu einer besseren Vorhersage zukünftiger Bewertungen.

4. In New-User-Szenarien führt die Berücksichtigung theoriebasierter Nutzermodelle zu einer besseren Erkennung relevanter Aktionen.

Generierung der Trainingsdaten In beiden Tests wurden Leave-One-Out-Kreuzvalidierungen durchgeführt. Diese sind ein Spezialfall der k-fachen Kreuzvalidierung, bei der der komplette Datensatz in k etwa gleich große Teile zerlegt wird und in k Durchgängen jeweils einer der Teile zum Testen (Abgleich der vorhergesagten Bewertungen) und die restlichen Teile zum Trainieren (Vorhersage fehlender Bewertungen) des Systems genutzt werden. Am Ende der Validierung werden die Ergebnisse der Durchgänge gemittelt. Populäre Varianten der Kreuzvalidierung im Bereich Empfehlungssysteme sind Validierungen mit $k = 5$ und $k = 10$. Bei kleinen Datensätzen wird häufig aber auch eine Leave-One-Out-Kreuzvalidierung durchgeführt, um eine ausreichende Datenbasis für das Training zu erhalten. Bei der Leave-One-Out-Kreuzvalidierung ist k gleichgesetzt mit der Anzahl der Nutzer im System. Es wird also in jedem Durchgang der Validierung ein Teil der Daten eines einzelnen Nutzers zum Testen verwendet.

Die genaue Zusammensetzung der Trainingsdatensätze hing in dieser Evaluation vom jeweils erwünschten Grad der Spärlichkeit (SG) der Daten ab.

$$SG = 1 - \frac{|Bewertungen|}{|Nutzer| * |Objekte|} \tag{4.4}$$

Dementsprechend berechnete sich die Gesamtanzahl der Bewertungen in den Trainingsdaten durch:

$$|Bewertungen| = (1 - SG) * |Nutzer| * |Objekte| \tag{4.5}$$

Um realistische Trainingsdaten zu erhalten wurden zur Erstellung des Datensatzes aus dem Gesamtdatensatz zufällig, aber (in etwa) gleichmäßig über alle Nutzer verteilt, solange Bewertungen aus dem Datenset entfernt bis die gewünschte Spärlichkeit erreicht wurde. Dabei wurde beachtet, dass für alle Nutzer mindestens eine Bewertung erhalten blieb. Um eine Beeinflussung der Ergebnisse durch die Auswahl

der Daten auszuschließen, wurden der Sparsity-Test und der New-User-Test jeweils zehn Mal mit neu generierten Trainingsdaten wiederholt. Bei der anschließenden Auswertung wurden die durchschnittlichen Ergebnisse der Durchläufe berechnet. Des Weiteren wurden die Tests mit einer variierenden Anzahl k an Nachbarnutzern, die zur Vorhersage der Bewertungen berücksichtigt wurden, wiederholt. Die minimale Anzahl war $k = 5$. Sie steigerte sich in 5er-Schritten bis hin zu allen vorhandenen Nutzern ($k_{max}(CARE) = 50$; $k_{max}(SavER) = 89$). Für die Analyse der Performanz der Filteralgorithmen wurde, analog zur Arbeit von Hu und Pu [Hu und Pu, 2011], jeweils das beste Ergebnis dieser Wiederholungen herangezogen.

4.3.2 Evaluation im Anwendungsszenario CARE

Generierung eines Datensatzes Für die Datensammlung im Anwendungsszenario CARE wurde ein zweiteiliger Fragebogen erstellt. Den ersten Teil des Fragebogens stellte die deutsche Version des WHOQOL-BREF Fragebogens [Skevington et al., 2004] dar. Im zweiten Teil des Fragebogens wurden die Befragten mit insgesamt 31 Empfehlungen zur Verbesserung des Wohlbefindens konfrontiert. Diese stammten allesamt aus der Datenbank des CARE-Projekts. Die genaue Verteilung der Aktionen auf die verschiedenen Kategorien von Wohlbefinden war: körperliches Wohlbefinden - körperliche Übungen (8), körperliches Wohlbefinden - Ernährung (6), mentales Wohlbefinden (4), soziales Wohlbefinden (4), emotionales Wohlbefinden (7), Umfeld/Umgebung (2). Die Teilnehmer der Befragung sollten alle aufgelisteten Aktionen auf einer Skala von 1 = “Gefällt mir gar nicht“ bis 5 = „Gefällt mir sehr gut“ bewerten.

Analyse des Datensatzes Der Fragebogen wurde ausgedruckt und digital verteilt. Insgesamt nahmen an der Befragung 37 Frauen und 16 Männer im Alter zwischen 50 und 93 Jahren (Durchschnitt: 74,5) teil. Wie in Abbildung 4.9 zu sehen ist, wurden nicht nur alleinstehende Senioren befragt. Das Hauptziel von CARE ist zwar die Unterstützung alleinstehender älterer Menschen, das System sollte aber auch allen anderen Senioren dabei helfen können, ihr Wohlbefinden zu fördern.

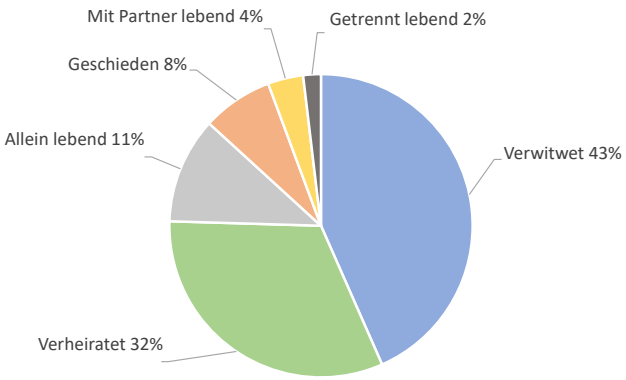

Abbildung 4.9: Familienstände der Teilnehmer der Datensammlung

Der Anleitung des WHOQOL-BREF folgend wurde bei einzelnen fehlenden Antworten in einer Kategorie des Fragebogens die durchschnittliche Bewertung für die restlichen Fragen dieser Kategorie als Ersatz verwendet. Dieses Verfahren wurde auch für die Bewertungen der Empfehlungen übernommen. Zwei Teilnehmer gaben allerdings für eine zu große Anzahl der Fragen keine Bewertung ab, so dass ihre Daten komplett entfernt werden mussten. Somit konnten lediglich die Daten von 51 der Befragten für die Evaluation verwendet werden.

Ermittlung geeigneter Parameterwerte für die hybriden Algorithmen Da die Qualität der hybriden Verfahren von der Wahl ihrer Parameter abhängt, wurden vor der eigentlichen Evaluation in Testläufen gut geeignete Parameter ermittelt. Hierfür wurden die beiden hybriden Filter mit verschiedenen Parametereinstellungen mit Graden der Spärlichkeit in 5%-Schritten von 25% bis 95% getestet. Als Evaluationsmetriken wurden der MAE und das F1-Maß herangezogen.

Beim linearen Ähnlichkeitsmaß wird mittels eines α die Gewichtung der bewertungsbasierten und theoriebasierten Ähnlichkeiten gesteuert. Hu und Pu [Hu und Pu, 2011] wählten ein dynamisches α von $\rho * |I_u \cap I_v| / (|I_u \cap I_v| + \delta)$, um die Gewichtung an die Anzahl der Bewertungen im System anpassen zu können. Je kleiner ρ gewählt wird, umso stärker ist der Einfluss der theoriebasierten Ähnlichkeit. Über δ kann gesteuert werden, ob mehr oder weniger Bewertungen vorhanden sein müssen, ehe das Gewicht der bewertungsbasierten Ähnlichkeit steigt. Hu und Pu wählten in ihrer Arbeit $\rho = 0.8$ und $\delta = 5$. Da ein dynamisches α auch für die eigene Evaluation als sinnvoll erachtet wurde, wurde Hus und Pus Ansatz übernommen. Allerdings wurden die Parameter ρ und δ nochmals in Testläufen mit $0,1 \leq \rho \leq 0,9$ und δs von 3, 5 und 10 evaluiert. Die Ergebnisse zeigten, dass mit $\rho = 0,9$ und $\delta = 3$ das beste Ergebnis erzielt werden konnte.

Der Ansatz mit Merkmalserweiterung kann durch die Anzahl der Nachbarn β beeinflusst werden, die zur Generierung der „virtuellen“ Bewertungen berücksichtigt werden. Zur Bestimmung des besten Wertes für diesen Parameter wurde die Performanz des Verfahrens für fünf, zehn, 15, 20 und 25 Nachbarn verglichen. Es zeigte sich, dass mit $\beta = 25$ die besten Ergebnisse erzielt werden konnten.

Evaluation in einem Sparsity-Szenario Im Sparsity-Szenario wurde die Performanz der Filteransätze anhand von Trainingssets mit Graden der Spärlichkeit (SG) von 25% bis 95% (5%-Schritte) untersucht. Niedrigere Spärlichkeit wurde nicht untersucht, da in diesen unrealistischeren Fällen von allen Nutzern bereits für beinahe alle Objekte Bewertungen vorliegen würden, siehe Tabelle 4.1.

Ergebnisse - Vorhersage der Bewertungen In Abbildung 4.10 sind bzgl. der Entwicklung des MAE mit steigender Spärlichkeit zwei Phasen zu erkennen.

Bei einer Spärlichkeit von 25% bis ca. 55%, also mit vielen vorhandenen Bewertungen, schnitten alle Filterverfahren in etwa gleich ab. Die Unterschiede rangierten meistens zwischen 0% und 4%, siehe Anhang A.1. Speziell bei einer sehr geringen

Tabelle 4.1: Beispielhafte Größe der Trainingsdatensätze abhängig vom Grad der Spärlichkeit (SG) (Nutzer: 51, Empfehlungen: 31, Mögliche Bewertungen: 1581)

SG	Anzahl Bewertungen	
	Trainingsset	∅ pro Nutzer
5	1502	29
25	1186	23
50	791	16
75	395	8
95	79	2

Spärlichkeit (SG≤40%) machte sich aber beim rein auf dem Wohlbefindens-Modell basierenden Verfahren die fehlende Berücksichtigung der Bewertungen mit einem 5-7% schlechteren MAE bemerkbar. Ein Friedman-Test mit anschließendem Dunn-Bonferroni-Post-Hoc-Test bestätigte, dass das theoriebasierte Verfahren in der Phase bis SG= 55% sogar signifikant schlechter Bewertungen vorhersagte, als das bewertungsbasierte und linear hybride Verfahren, siehe Tabelle 4.2.

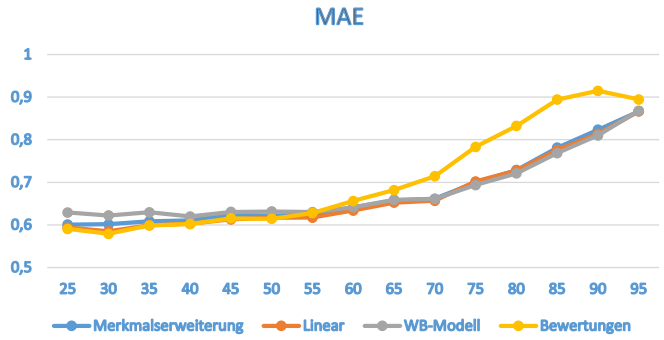

Abbildung 4.10: Qualität der Vorhersage der Bewertungen im Sparsity-Szenario

Je weniger Bewertungen im System vorhanden waren ($SG > 60\%$), desto größer war der MAE beim bewertungsbasierten Ansatz. Die neuen Verfahren, die das Wohlbefinden der Nutzer bei der Empfehlungsauswahl berücksichtigen, konnten diesen Mangel an Bewertungen besser kompensieren, so dass der MAE dieser Verfahren weniger groß ausfiel. Besonders stark war der Unterschied bei einer Spärlichkeit zwischen 75% und 90%. Nur wenn kaum Bewertungen vorhanden waren ($SG = 95\%$), glich sich der MAE aller Verfahren wiederum an. Auch diese Erkenntnisse konnten durch die genannten Signifikanztests belegt werden. Für eine Spärlichkeit zwischen 60% und 95% schnitten sowohl der lineare hybride Ansatz als auch der theoriebasierte Ansatz signifikant besser ab, als der bewertungsbasierte Ansatz, siehe Tabelle 4.2.

Da sich die lineare Kombination in beiden Phasen als eines der besten Verfahren herauskristallisierte, wurden die Signifikanztests auch über alle Grade der Spärlichkeit hinweg durchgeführt. Der Friedman-Test deutete zwar auf signifikante Unterschiede hin und das lineare Filterverfahren erreichte das beste Ergebnis. Der paarweise Vergleich ergab aber keine signifikanten Unterschiede.

Tabelle 4.2: Ergebnisse der Signifikanztests für den MAE im Sparsity-Szenario (Abkürzungen: SG = Grad der Spärlichkeit; BB = Bewertungsbasierter kollaborativer Filter; WB = Theoriebasierter Filter mit Wohlbefindens-Modell; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter)

Verfahren	Mittlerer Rang (MAE)		
	SG≤55%	SG>55%	25%≤SG≤95%
BB	1,50	4,00	2,83
WB	4,00	1,88	2,87
ME	2,86	2,44	2,63
LIN	1,64	1,69	1,67
Friedman	$\chi^2(3) = 17,522$**	$\chi^2(3) = 16,481$**	$\chi^2(3) = 8,856$*
Post-Hoc (Dunn-Bonferroni)	WB<BB** (z=3,623) WB<LIN** (z=-3,416)	BB<LIN** (z=-3,583) BB<WB** (z=-3,292)	
(*signifikant mit p < .05; **signifikant mit p < .01; ***signifikant mit p < .001)			

Ergebnisse - Klassifikation nach Relevanz Für die Klassifizierung der Objekte hinsichtlich ihrer Relevanz für die Nutzer fielen die Unterschiede zwischen den Filterverfahren weniger deutlich aus als beim MAE, siehe Abbildung 4.11. Bis zu einer Spärlichkeit von 75-80% lieferten, bis auf das rein theoriebasierte Verfahren, alle Verfahren in etwa gleich gute Ergebnisse, die sich um maximal 2% unterschieden, siehe Anhang A.1. Der theoriebasierte Ansatz schnitt speziell bei sehr vielen Bewertungen im System um bis zu 10% schlechter ab. Dieses Ergebnis spiegelte sich sowohl in der Precision als auch im Recall wider. Die Qualität der Klassifizierung des theoriebasierten Ansatzes war demzufolge im Bereich von 25%-80% auch signifikant schlechter als alle anderen Ansätze. Dies ergaben wiederum ein Friedman-Test und ein Post-Hoc-Test, deren Ergebnisse in Tabelle 4.3 aufgelistet sind.

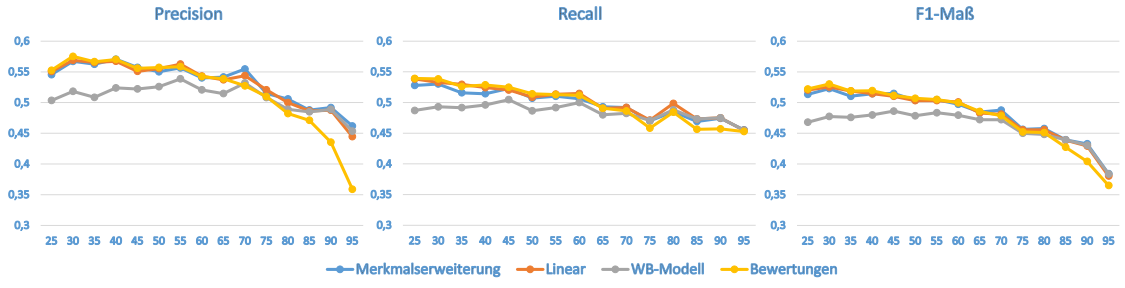

Abbildung 4.11: Qualität der Klassifizierungen nach Relevanz im Sparsity-Szenario

Bei einer Spärlichkeit größer als 80% fiel die Qualität des bewertungsbasierten Ansatzes stark ab. Das F1-Maß und speziell die Precision brachen erheblich ein, während die anderen Ansätze dank des zusätzlichen Wissens die Qualität höher halten konnten. Der Ansatz mit Merkmalserweiterung erreichte im Vergleich zum

bewertungsbasierten Ansatz sogar signifikant bessere Ergebnisse. Über alle Grade der Spärlichkeit hinweg schnitten der Hybrid mit Merkmalserweiterung und der bewertungsbasierte Filtertechnik signifikant besser ab als die theoriebasierte Technik. Die genauen Ergebnisse beider Signifikanztests sind in Tabelle 4.3 nachzulesen.

Tabelle 4.3: Ergebnisse der Signifikanztests für das F1-Maß im Sparsity-Szenario (Abkürzungen: SG = Grad der Spärlichkeit; BB = Bewertungsbasierter kollaborativer Filter; WB = Theoriebasierter Filter mit Wohlbefindens-Modell; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter)

Verfahren	Mittlerer Rang (F1-Maß)		
	SG≤80%	SG>80%	25%≤SG≤95%
BB	1,67	4,00	2,13
WB	4,00	2,00	3,60
ME	2,08	1,00	1,87
LIN	2,25	3,00	2,40
Friedman	$\chi^2(3) = 22,900^{***}$	$\chi^2(3) = 9,000^{*}$	$\chi^2(3) = 15,800^{**}$
Post-Hoc (Dunn-Bonferroni)	WB<BB*** (z=-4,427) WB<ME** (z=3,637) WB<LIN** (z=3,320)	BB<ME* (z=2,846)	WB<BB* (z=-3,111) WB<ME** (z=3,677)
(*signifikant mit p < .05; **signifikant mit p < .01; ***signifikant mit p < .001)			

Evaluation in einem New-User-Szenario Anhand des New-User-Szenarios wurde verglichen, wie gut die Filterverfahren in einem System, das bereits eine gewisse Menge an Bewertungen enthält (SG = 50%), mit neuen Nutzern umgehen können. Dem Testnutzer des jeweiligen Durchgangs des Leave-One-Out-Verfahrens wurden hierfür im Trainingsset zwischen einer und fünf Bewertungen zugewiesen.

Ergebnisse - Vorhersage der Bewertungen Abbildung 4.12 zeigt, dass der bewertungsbasierte Ansatz erwartungsgemäß größere Probleme mit der Vorhersage von Bewertungen für neue Nutzer hatte als die anderen Verfahren, die mehr Wissen über die Nutzer zur Verfügung hatten und fast identische Ergebnisse erzielten. Erst ab einer Menge von vier Bewertungen erreichte das bewertungsbasierte Verfahren ein ähnliches Niveau, siehe auch Anhang A.2. Ein Friedman-Test und ein Dunn-Bonferroni-Post-Hoc-Test bestätigten, dass die bewertungsbasierte Filtertechnik signifikant schlechter abschnitt, als der theoriebasierte Ansatz, der die besten Ergebnisse erreichte, siehe Tabelle 4.4.

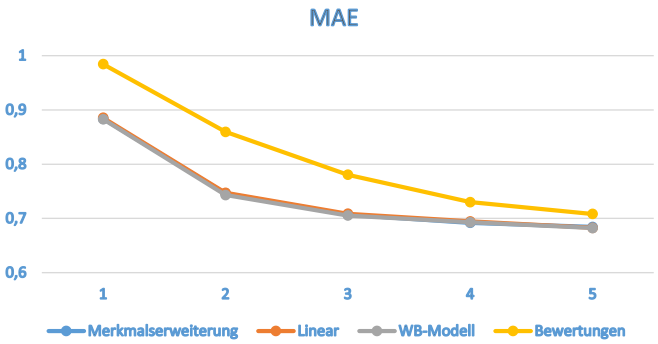

Abbildung 4.12: Qualität der Vorhersage der Bewertungen im New-User-Szenario

Ergebnisse - Klassifikation nach Relevanz Für die Einschätzung der Relevanz der Empfehlungen für die Nutzer zeigte sich auch im New-User-Szenario, dass das bewertungsbasierte Filterverfahren gerade im Hinblick auf die Precision Probleme hatte, siehe Abbildung 4.13. Dies schlug sich auch im F1-Maß nieder. Bis zu einer Menge von drei Bewertungen schnitten die anderen Verfahren besser ab. Laut eines Friedman-Tests erreichte das lineare hybride Filterverfahren insgesamt die besten Ergebnisse, siehe Tabelle 4.4. Allerdings gab es keine signifikanten Unterschiede.

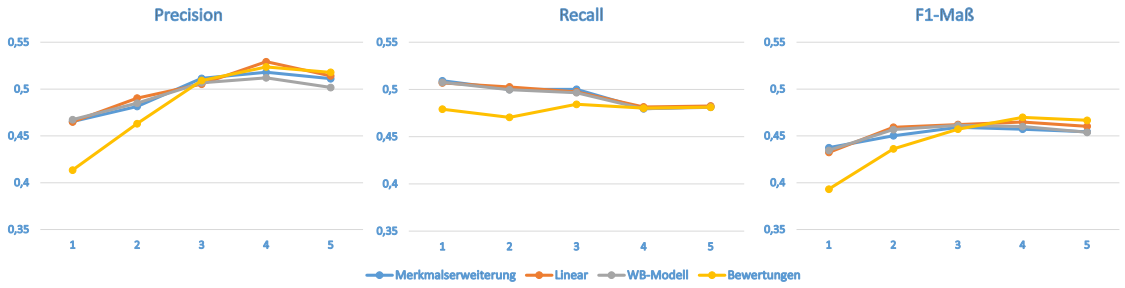

Abbildung 4.13: Qualität der Klassifizierungen nach Relevanz im New-User-Szenario

Fazit der Evaluation Die Evaluation der verschiedenen Ansätze zur kollaborativen Empfehlungsauswahl im CARE-Szenario hat gezeigt, dass die Berücksichtigung des Wissens über das aktuelle Wohlbefinden der Nutzer sich speziell in Cold-Start-Szenarien vorteilhaft auf die Qualität der Empfehlungen auswirkt. Dies betrifft vor allem die Qualität der vorhergesagten Bewertungen, so dass die Hypothesen 1 und 3 bestätigt werden konnten. Für die Klassifikation der Objekte hinsichtlich ihrer Relevanz für die Nutzer waren die Unterschiede weniger deutlich. Die Hypothesen 2 und 4 konnten jeweils nur für Situationen mit sehr wenigen Bewertungen bestätigt werden. Die Evaluation zeigte allerdings auch, dass das rein theoriebasierte Verfahren speziell in Situationen mit vielen Bewertungen ebenfalls Schwächen hat. Am besten und in etwa gleich schnitten die beiden hybriden Filterverfahren ab.

Tabelle 4.4: Ergebnisse der Signifikanztests im New-User-Szenario (Abkürzungen: BB = Bewertungsbasierter kollaborativer Filter; WB = Theoriebasierter Filter mit Wohlbefindens-Modell; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter)

Verfahren	Mittlerer Rang	
	MAE	F1-Maß
BB	4,0	2,8
WB	1,2	2,6
ME	2,0	2,8
LIN	2,8	1,8
Friedman	$\chi^2(3) = 12,840$**	$\chi^2(3) = 2,040$; $p = 0,564$
Dunn-Bonferroni	BB<WB** (z=3,429)	
(*signifikant mit p < .05; **signifikant mit p < .01; ***signifikant mit p < .001)		

4.3.3 Evaluation im Anwendungsszenario SavER

Um die Ergebnisse der Untersuchungen im CARE-Szenario bestätigen und den generellen Vorteil der Berücksichtigung theoriebasierter Nutzermodelle für die Empfehlungsauswahl in beratenden Empfehlungssystemen belegen zu können, wurden die alternativen kollaborativen Filterverfahren auch im SavER-Szenario evaluiert.

Generierung eines Datensatzes [Hammer et al., 2015a] Der Online-Fragebogen für die Datensammlung bestand wiederum aus zwei Teilen.

Der erste Teil zielte auf eine Einschätzung der Energiekultur der Befragten ab. Da hierfür kein ausreichend kurzer, validierter Fragebogen zur Verfügung stand, wurde ein eigener Fragebogen erstellt. Dieser fragte charakteristische Eigenschaften und Verhaltensweisen der in Kapitel 4.2.2 beschriebenen Energiekulturen ab. Die eingesetzten Likert-Skalen umfassten die Optionen von 1=„auf keinen Fall“ bis 5=„auf jeden Fall“. Zur Einschätzung der *Normen* sollten Aussagen wie „Ich mache mir keine Gedanken über meinen Energieverbrauch.“, „Ich finde Energiesparen wichtig, unternehme aber selbst noch zu wenig.“ oder „Ich achte sehr auf meinen Energieverbrauch.“ bewertet werden. Außerdem wurde berücksichtigt, ob der Schutz der Umwelt für die Teilnehmer einen Grund für energiesparendes Verhalten darstellte. Im Bereich *materielle Kultur* sollte u.a. die Hausisolierung bewertet werden. Zusätzlich sollten Angaben zur eingesetzten Heizungsart und zur Anzahl vorhandener Haushaltsgeräte, motorisierter Fahrzeuge und Fahrräder gemacht werden. Um eine Pro-Kopf-Anzahl der Geräte und Fortbewegungsmittel berechnen zu können, wurde die Anzahl der Personen im Haushalt abgefragt. Wichtig zu erwähnen ist, dass nicht nur die Anzahl der Geräte und Fortbewegungsmittel bewertet wurde, sondern ihre Auswirkung auf den Energieverbrauch. Das tatsächliche *Energieverhalten* der Nutzer wurde mit Hilfe ihrer Bewertungen im zweiten Teil des Fragebogens eingeschätzt.

In diesem Teil des Fragebogens waren 21 Energiespartipps aus den Bereichen Heizenergie, Stromverbrauch und Benzinverbrauch aufgelistet, siehe Tabelle 4.5. Die Liste beruhte u.a. auf Energiespartipps der WWF[15] und des Bundesministeriums für Umwelt[16]. Bei der Auswahl der Energiespartipps wurde darauf geachtet, dass ähnlich viele Tipps mit unterschiedlich hohem finanziellen, körperlichen oder zeitlichen Aufwand vorhanden waren. Es gab Empfehlungen zu baulichen Veränderungen am Gebäude, zur Anschaffung energieeffizienterer Geräte und Fahrzeuge sowie zu Verhaltensänderungen. Beispiele sind „Sie könnten Energie sparen, wenn Sie...“

„...Geräte komplett ausschalten anstatt die Stand-By-Funktion zu verwenden.“

„...Ihre Wäsche für das Trocknen aufhängen anstatt den Trockner zu nutzen.“

„...Ihre bisherigen Leuchtmittel durch energiesparende Leuchtmittel ersetzen.“

Die Energiespartipps sollten zweimal bewertet werden. Zunächst sollte das Interesse gegenüber der jeweiligen Aktion auf einer Skala von 1 = „überhaupt nicht interessant“ bis 5 = „sehr interessant“ angegeben werden. Danach wurden die Teilnehmer nach ihrer Bereitschaft zur Ausführung der Aktionen befragt. Die hierfür verwendeten semantischen Differenziale der 5er-Likert-Skala lauteten „würde ich nicht umsetzen“ und „würde ich sicher umsetzen“. Zusätzlich gab es die Option „mache/besitze ich bereits“. Die Anzahl der mit dieser Option bewerteten Energiesparaktionen wurde zur Einschätzung des aktuellen *Energieverhaltens* genutzt.

Analyse des Datensatzes [Hammer et al., 2015a] An der Umfrage nahmen 32 Frauen und 58 Männer teil, die einen guten Querschnitt der Zielgruppe des SavER-Systems darstellten. Zu ihnen zählten junge, unverheiratete Studenten mit geringem Einkommen, die in Städten in Mietwohnungen oder Wohngemeinschaften wohnten. Menschen mittleren Alters mit verschiedensten Familienständen und durchschnittlichem Einkommen, die zum Teil in ihren Eigenheimen in ländlichen Regionen oder aber auch in Großstädten zur Miete wohnten, gehörten ebenfalls zu den Teilnehmern. Die dritte große Teilnehmergruppe setzte sich aus Menschen zusammen, die älter als 50 Jahre, meist verheiratet und Eltern erwachsener Kinder waren. Sie hatten durchschnittliche bis hohe Einkommen und wohnten meist in kleineren Städten oder ländlichen Gebieten in eigenen oder gemieteten Häusern.

Die Teilnehmer wurden hinsichtlich ihrer materiellen Kultur, ihrer Normen und ihres Energieverhaltens bewertet und anschließend gruppiert. Wie bereits erwähnt, gab es von Seiten des Energy Cultures Framework keinerlei Vorgaben, wie die drei Faktoren genau charakterisiert oder gemessen werden. Nach einer ausführlichen Analyse der gesammelten Daten wurde folgendes System zur Gruppierung der Nutzer angewandt: Für jede Person wurde für jeden der drei Kulturdimensionen der Anteil an Antworten berechnet, der jeweils hinsichtlich der Umweltbilanz als positiv

[15] http://www.wwf.de/aktiv-werden/tipps-fuer-den-alltag/energie-spartipps/strom-sparen/

[16] http://www.bmub.bund.de/themen/klima-energie/energieeffizienz/foerdermittel-beratung/energiespartipps/

Tabelle 4.5: Energiespartipps gruppiert nach den Ergebnissen der Befragung

(1)

- Stoßlüften anstatt dauerhaft gekippter Fenster
- Herunterregeln der Heizung in ungenutzten Räumen
- Ausschalten des Lichts in ungenutzten Räumen

(2)

- Heizkörper regelmäßig entlüften
- Erhalt einer konstanten Raumtemperatur durch Schließen der Rollläden oder Vorhänge
- Kauf energieeffizienterer Lampen
- Trocknen der Wäsche im Freien
- Kühlschrank auf 7° einstellen
- Vermeidung des Stand-By Modus elektrischer Geräte
- Häufigeres Zufußgehen/Radfahren

(3)

- Kauf von Zugluftstoppern für Türen und Fenster
- Erneuern der Fenster
- Installation intelligenter oder programmierbarer Thermostate
- Kauf energieeffizienterer Geräte
- Ausstecken ungenutzter Geräte
- Häufigere Nutzung des ÖPNV

(4)

- Verbesserung der Hausisolierung
- Anbringen Wärme reflektierender Matten hinter Heizkörpern
- Erneuerung des Heizsystems
- Kauf eines Hybrid- oder Elektroautos
- Teilnahme an einem Car Sharing-Programm

bewertet werden konnte. War der Anteil an positiven Bewertungen für eine Kulturdimension größer als 60% führt dies zu einer insgesamt positiven Bewertung der Dimension. Anteile zwischen 40% und 60% führten zu einer neutralen Bewertung. Lag der Anteil der positiven Bewertungen unterhalb 40% so wurde die Dimension negativ bewertet.

Ein Beispiel für die Bewertung der materiellen Kultur soll das Bewertungssystem noch einmal auf einfache Weise erklären. Für die materielle Kultur wurden insgesamt sechs Faktoren berücksichtigt: (1) subjektive Einschätzung der Gebäudeisolierung, (2) Heizungstyp sowie der Pro-Kopf-Anteil der im Haushalt vorhandenen (3) Elektrogeräte, (4) Autos, (5) motorisierten Zweiräder und (6) Fahrräder. Für jeden dieser Faktoren wurde eine positive, neutrale oder negative Bewertung im Bezug auf den Energiebilanz getroffen. Die Einschätzung der Heizarten beruhte auf einer Recherche auf verschiedenen Ratgeber-Seiten [Heizsparer.de, 2017, Strom-Magazin.de, 2017].

Betrachten wir beispielsweise einen Studenten, der in einer Altbauwohnung mit schlechter Isolierung (negativ) und einer Ölheizung (negativ) wohnt. Er besitzt eine durchschnittliche Anzahl an Geräten (neutral) und weder ein Auto (positiv), noch ein motorisiertes Zweirad (positiv). Allerdings besitzt er ein Fahrrad (positiv). Er erhält also in 50% der Bewertungen eine positives Bewertung, was zu einer neutralen Einschätzung hinsichtlich seiner materiellen Kultur führt.

Die Analyse aller Befragten zeigte die Verteilung, wie sie in Abbildung 4.14 zu sehen ist. Ein großer Teil der Befragten hatte bereits eine positive Einstellung zum Energiesparen und die Hälfte der Teilnehmer zeichnete sich auch durch energiesparendes Verhalten aus. Lediglich bei der materiellen Kultur fielen die Bewertungen etwas schlechter aus, was u.a. daran lag, dass die jüngeren Teilnehmer häufig in älteren Gebäuden mit schlechter Isolation und älteren Heizsystemen wohnten.

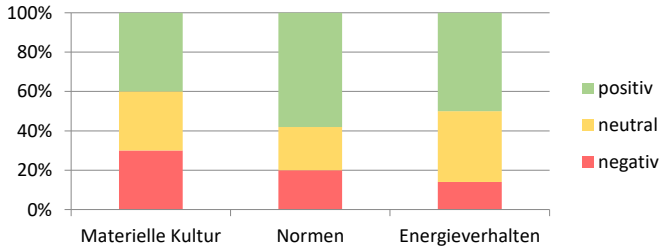

Abbildung 4.14: Einschätzung der Teilnehmer hinsichtlich ihrer materiellen Kultur, Normen und ihres tatsächlichen Energieverhaltens

Die durchschnittlichen Bewertungen der Energiespartipps durch die Befragten ergab vier Kategorien von Empfehlungen, siehe Tabelle 4.5:

1. Die erste Kategorie setzte sich aus häufig bekannten Aktionen zusammen, die von den meisten Personen bereits ausgeführt wurden. Sie verursachen kaum Aufwand und sind für die meisten Leute selbstverständlich.

2. Etwas weniger bekannt waren die Aktionen der zweiten Kategorie. Etwas die Hälfte der Teilnehmer führte auch diese Aktionen bereits aus und ein großer Teil der restlichen Nutzer (40-70%) hielt die Aktionen für interessant und zeigte Bereitschaft sie auch auszuführen. Die Aktionen dieser Kategorie sind bereits weniger selbstverständlich als die Aktionen der ersten Kategorie und verursachen in einem geringen Maße auch bereits körperlichen, zeitlichen oder auch finanziellen Aufwand.

3. Über die Aktionen der dritten Kategorie herrschte Uneinigkeit unter den Befragten. Jeweils ungefähr ein Drittel der Personen führt die Aktionen bereits aus, zeigte Bereitschaft sie auszuführen oder zeigte kein Interesse an den Aktionen. Im Vergleich zur zweiten Kategorie bedeuten die Aktionen einen stärkeren Aufwand und sind, wie zum Beispiel im Falle der Erneuerung der Fenster oder der regelmäßigen Nutzung des ÖPNV, nicht mehr für alle Personen möglich.

4. Aktuell noch zu unbekannte oder zu exotische Aktionen wie reflektierende Matten hinter den Heizungen oder Car Sharing und für viele Teilnehmer nicht durchführbare Aktionen bildeten die letzte Kategorie. Zu ihnen gehörten bauliche Änderungen sowie Anschaffungen, die einen hohen finanziellen Aufwand mit sich bringen. Nur ein kleiner Teil der Teilnehmer ($< 20\%$) führte diese Aktionen bereits aus und die meisten der anderen Teilnehmer zeigten weder Interesse noch Bereitschaft an der Ausführung dieser Energiespartipps.

Um einen ersten Eindruck davon zu erhalten, ob sich energiekulturelle Hintergründe auf die Präferenzen für Energiespartipps auswirken, wurde eine deskriptive Analyse der durchschnittlichen Bewertungen durchgeführt, vgl. Abbildung 4.15.

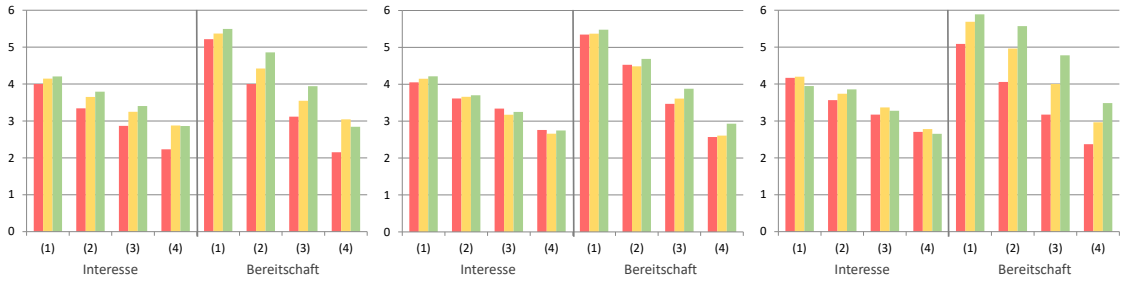

Abbildung 4.15: Durchschnittliche Bewertungen der Energiespartipps der vier Kategorien. Abgegeben von Teilnehmern mit einer positiven (grün), neutralen (gelb) und negativen (rot) Einschätzung für die Faktoren Normen (Links), materielle Kultur (Mitte), Energieverhalten (rechts)

Die Analyse zeigte, dass unterschiedliche Normen durchaus zu unterschiedlichen Präferenzen führten. Speziell bei weniger beliebten Energiespartipps zeigten Personen mit einer positiven Einstellung zum Thema Energiesparen mehr Interesse und Bereitschaft zur Durchführung der Maßnahmen. Die materielle Kultur hatte hingegen nur einen geringen Einfluss auf die Präferenzen der Teilnehmern. Personen mit einer positiven Bewertung für diesen Faktor zeigten lediglich eine etwas erhöhte Bereitschaft zur Durchführung der präsentierten Aktionen. Da die Bereitschaft Energiespartipps auszuführen das hauptsächliche Kriterium für die Einschätzung des aktuellen Energieverhaltens war, konnte für diese Kategorisierung nur das Interesse der Teilnehmer analysiert werden. Interessanterweise zeigten sich hier kaum Unterschiede zwischen den einzelnen Nutzergruppen. Energiesparsame Teilnehmer zeigten teilweise sogar weniger Interesse, als die anderen Teilnehmer. Dies lässt sich u.a. dadurch begründen, dass engagierte Personen viele der präsentierten Aktionen bereits kannten oder ausführten.

Zusammenfassend bestätigte diese erste Analyse der gesammelten Daten die Erkenntnisse aus der Verhaltenspsychologie, siehe Kapitel 2.1, und damit auch die Sinnhaftigkeit das Konzept der Energiekultur bei die Empfehlungsauswahl des SavER-Systems miteinzubeziehen. Die Daten ließen zum Beispiel darauf schließen, dass Menschen, die sich bereits durch energiebewusstes Verhalten auszeichnen, kein Interesse mehr an altbekannten Energiespartipps haben, sondern eher auf der Suche

nach neuen, möglicherweise auch aufwendigeren Aktionen sind. Im Gegensatz dazu sollten Personen, die sich bisher nur wenige Gedanken über ihren Energieverbrauch gemacht haben, nicht mit aufwendigen Aktionen überfordert werden. Diese Personengruppe sollte zunächst durch Empfehlungen für einfache Maßnahmen an das Thema heran geführt und von der Notwendigkeit und den Vorteilen energiesparenden Verhaltens überzeugt werden.

Ermittlung geeigneter Parameterwerte für die hybriden Algorithmen Um auch im SavER-Szenario für die hybriden Filtertechniken bestmögliche Ergebnisse erreichen zu können, wurden analog zum CARE-Szenario, siehe Kapitel 4.3.2, in Testläufen die besten Parameterwerte für die beiden Verfahren ermittelt. Für das lineare Ähnlichkeitsmaß wurden die besten Ergebnisse mit den Parametereinstellungen $\rho = 0,7$ (Gewichtung des grundsätzlichen Einflusses des theoriebasierten Ansatzes) und $\delta = 5$ (Richtwert für eine adäquate minimale Anzahl an Bewertungen) erzielt. Beim Ansatz mit Merkmalserweiterung führte wie im CARE-Szenario ein β von 25 (Anzahl der Nachbarn zur Generierung der "virtuellen" Bewertungen) zu den besten Ergebnissen.

Evaluation in einem Sparsity-Szenario Die Performanz der Filteransätze wurde mit Graden der Spärlichkeit (SG) von 25% bis 95% (5%-Schritte) evaluiert. Tabelle 4.6 zeigt Beispiele für die Anzahl an Bewertungen je Grad der Spärlichkeit.

Tabelle 4.6: Beispielhafte Größe der Trainingsdatensätze abhängig vom Grad der Spärlichkeit (SG) (Nutzer: 90, Empfehlungen: 21, Mögliche Bewertungen: 1890)

SG	Anzahl Bewertungen	
	Trainingsset	∅ pro Nutzer
25	1418	16
50	945	11
75	473	5
95	95	1

Ergebnisse - Vorhersage der Bewertungen Bei einer größeren Menge an Bewertungen im System (bis 55%-60% Spärlichkeit) ergab sich für alle Algorithmen ein MAE auf relativ ähnlichem Niveau, siehe Abbildung 4.16. Lediglich der rein theoriebasierte Ansatz schnitt hier schlechter ab. Bei einer Spärlichkeit größer als 60% verschlechterte sich die Qualität der vorhergesagten Bewertungen beim bewertungsbasierten Ansatz stark. Die neuen Ansätze konnten dagegen ihr Niveau bis zu einem gewissen Punkt (ca. 80%) stabiler halten. Bei 80-85% war der Qualitätsgewinn der Verfahren, die ein theoriebasiertes Nutzermodell einsetzen, gegenüber der klassischen Variante mit über 10% besonders groß, siehe Anhang A.3.

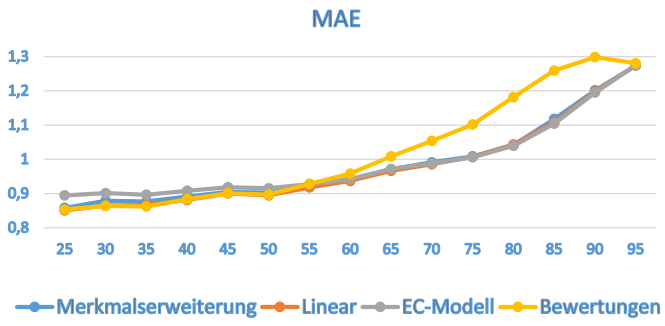

Abbildung 4.16: Qualität der Vorhersage der Bewertungen im Sparsity-Szenario

Ergebnisse - Klassifikation nach Relevanz Ähnlich wie beim MAE verhielt es sich bei der Fähigkeit der Verfahren die Relevanz der Empfehlungen für die jeweiligen Nutzer einzuschätzen. Der Verlauf des F1-Maß in Abbildung 4.17 zeigt, dass der theoriebasierte Ansatz bei einer geringeren Spärlichkeit schwächer abschnitt. Wiederum bei ca. 55% änderte sich das Verhältnis. Lediglich der Ansatz mit Merkmalserweiterung übertraf die anderen Ansätze unabhängig von der Sparsity. Interessant ist, dass die schlechten Werte hinsichtlich des F1-Maß bei einer hohen Spärlichkeit allein durch eine schlechte Precision zu erklären scheinen. Der Recall blieb für alle Filteransätze auf einem konstanten, aber mittelmäßigem Niveau. Alle Verfahren fanden also relativ gut relevante Objekte. Mit steigender Spärlichkeit wurden aber auch immer mehr nicht relevante Objekte zur Empfehlung ausgewählt.

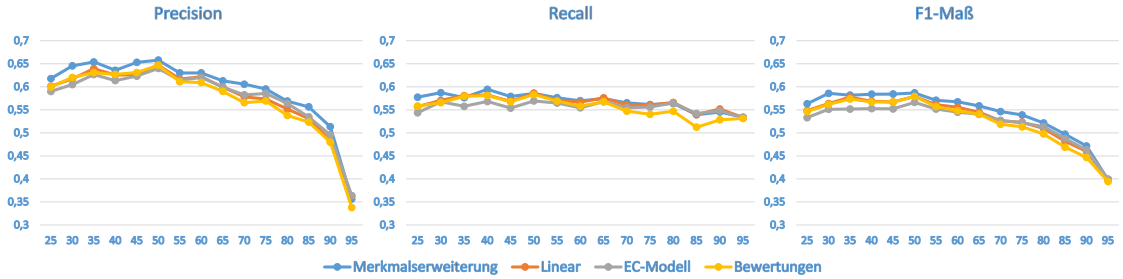

Abbildung 4.17: Qualität der Klassifizierungen nach Relevanz im Sparsity-Szenario

Ergebnisse - Signifikanztests Um die Erkenntnisse der deskriptiven Analyse zu bestätigen, wurden Signifikanztests nach Friedman und Dunn-Bonferroni-Post-Hoc-Tests durchgeführt.

Zunächst wurden die Phasen mit mehr (SG$\leq$55%) und weniger Bewertungen (SG>55%) getrennt voneinander untersucht, siehe Tabelle 4.7. In der Phase mit einer geringen Spärlichkeit erzielte der theoriebasierte Ansatz einen signifikant schlechteren MAE als der bewertungsbasierte und der lineare hybride Ansatz. Hinsichtlich des F1-Maßes schnitt er signifikant schlechter ab als die beiden hybriden Filterverfahren. Es bestätigte sich außerdem, dass der hybride Ansatz mit Merkmalserweiterung bereits in dieser Phase am besten dazu geeignet war die Relevanz von Empfehlungen

einzuschätzen. Er übertrumpfte nicht nur den theoriebasierten Ansatz signifikant, sondern auch den bewertungsbasierten Ansatz. Des Weiteren erzielte er auch für die höheren Grade der Spärlichkeit die besten Ergebnisse. Für diese Phase bestätigten die Signifikanztests auch, dass der bewertungsbasierte Ansatz am schlechtesten die Bewertungen vorhersagen und die Relevanz der Objekte einschätzen konnte. Im Vergleich zum linearen und theoriebasierten Verfahren (MAE) und zum Verfahren mit Merkmalserweiterung (F1-Maß) waren die Unterschiede sogar signifikant.

Tabelle 4.7: Ergebnisse der Signifikanztests im Sparsity-Szenario (Abkürzungen: SG = Grad der Spärlichkeit; BB = Bewertungsbasierter kollaborativer Filter; EC = Theoriebasierter Filter mit Energiekulturen; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter)

Verfahren	Mittlerer Rang			
	SG≤55%		SG>55%	
	MAE	F1-Maß	MAE	F1-Maß
BB	2,00	2,86	4,00	3,75
EC	3,86	4,0	1,88	2,50
ME	2,86	1,0	2,38	1,12
LIN	1,29	2,14	1,75	2,62
Friedman	$\chi^2(3) = 15,5^{**}$	$\chi^2(3) = 19,97^{***}$	$\chi^2(3) = 15,45^{**}$	$\chi^2(3) = 16,65^{**}$
Post-Hoc (Dunn-Bonferroni)	EC<LIN** (z=-3,73) EC<BB* (z=2,69)	EC<ME*** (z=4,35) EC<LIN* (z=2,69) BB<ME* (z=2,69)	BB<LIN** (z=-3,49) BB<EC** (z=-3,29)	BB<ME*** (z=4,067)
(*signifikant mit p < .05; **signifikant mit p < .01; ***signifikant mit p < .001)				

Da die bisherigen Ergebnisse darauf hindeuteten, dass der lineare hybride Ansatz unabhängig vom Grad der Sparsity am besten zur Vorhersage von Bewertungen geeignet ist und, dass der hybride Ansatz mit Merkmalserweiterung ebenfalls unabhängig vom Grad der Sparsity am besten die Relevanz von Objekten einschätzen kann, wurden die Signifikanztests nochmals mit den kompletten Daten (25%≤SG≤95%) durchgeführt. Die Ergebnisse in Tabelle 4.8 zeigen, dass sich diese Annahmen bestätigten. Der lineare hybride Filteransatz traf signifikant bessere Vorhersagen über die Bewertungen der Nutzer als die Ansätze, die sich allein auf die Bewertungen oder die Energiekultur der Nutzer stützten. Der Hybrid mit Merkmalserweiterung konnte dagegen die Empfehlungen signifikant besser hinsichtlich ihrer Relevanz für die Nutzer klassifizieren als alle anderen Verfahren.

Tabelle 4.8: Ergebnisse der Signifikanztests im Sparsity-Szenario mit 25%≤SG≤95% (Abkürzungen: BB = Bewertungsbasierter kollaborativer Filter; EC = Theoriebasierter Filter mit Energiekulturen; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter)

Verfahren	Mittlerer Rang	
	MAE	F1-Maß
BB	3,07	3,33
EC	2,80	3,20
ME	2,60	1,07
LIN	1,53	2,40
Friedman	$\chi^2(3) = 12,2^{**}$	$\chi^2(3) = 29,24^{***}$
Post-Hoc (Dunn-Bonferroni)	BB<LIN** (z=-3,25) EC<LIN* (z=-2,69)	BB<ME*** (z=4,81) EC<ME*** (z=4,53) LIN<ME* (z=2,83)
(*signifikant mit p < .05; **signifikant mit p < .01; ***signifikant mit p < .001)		

Evaluation in einem New-User-Szenario Analog zur Evaluation im CARE-Szenario wurde für diese Evaluation $SG = 50\%$ gewählt und für die Testnutzer eine bis fünf Bewertungen (1er-Schritte) in das Trainingsset aufgenommen.

Die deskripitive Analyse des MAE, siehe Abbildung 4.18, und des F1-Maßes, siehe Abbildung 4.19, zeigte, dass sich die Performanz der Filterverfahren bereits ab drei bzw. zwei vorhandenen Bewertungen stabilisierte. Ab diesem Zeitpunkt schnitten alle Verfahren in etwa gleich ab, siehe Tabelle 4.9. Bei weniger Bewertungen hatte vor allem der bewertungsbasierte Ansatz größere Probleme. Speziell, wenn nur eine Bewertung abgegeben wurde, schnitten die Verfahren, die die Energiekultur der Nutzer berücksichtigen, hinsichtlich beider Evaluationsmaße um 12% bis 13% besser ab, siehe Anhang A.4. Für diese Verfahren war der MAE beinahe gleich, während beim F1-Maß das linear hybride Filtern etwas besser zu funktionieren schien.

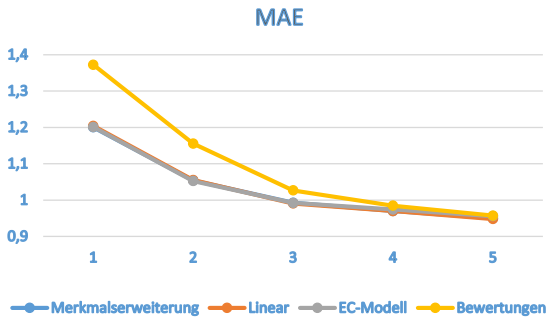

Abbildung 4.18: Qualität der Vorhersage der Bewertungen im New-User-Szenario

Ein Friedman-Test sowie ein Post-Hoc-Test nach Dunn-Bonferroni ergaben bzgl. der Qualität der Klassifikation nach der Relevanz keine signifikanten Unterschiede.

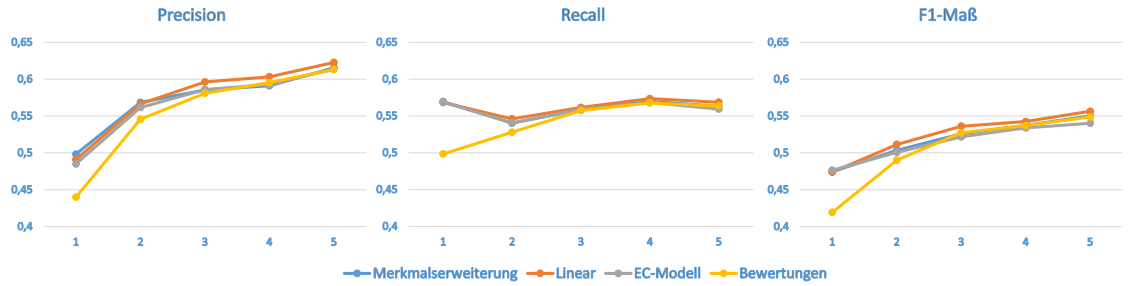

Abbildung 4.19: Qualität der Klassifizierungen nach Relevanz im New-User-Szenario

Der linear hybride Filter schnitt aber beinahe für alle Grade der Spärlichkeit am besten ab. Bei der Vorhersage der Bewertungen erreichte das rein bewertungsbasierte Filtern durchgängig die schlechtesten Werte und war signifikant schlechter als der hybride Filter mit Merkmalserweiterung, siehe Tabelle 4.9.

Tabelle 4.9: Ergebnisse der Signifikanztests im New-User-Szenario (Abkürzungen: BB = Bewertungsbasierter kollaborativer Filter; EC = Theoriebasierter Filter mit Energiekulturen; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter)

Verfahren	Mittlerer Rang	
	MAE	F1-Maß
BB	4,0	3,2
EC	2,2	3,2
ME	1,8	2,2
LIN	2,0	1,4
Friedman	$\chi^2(3) = 9,240$*	$\chi^2(3) = 6,840$; $p = 0,77$
Dunn-Bonferroni	BB<ME* (z=-2,694)	
(*signifikant mit p < .05; **signifikant mit p < .01; ***signifikant mit p < .001)		

Fazit der Evaluation Ähnlich wie im CARE-Szenario stellte auch im SavER-Szenario die Berücksichtigung eines theoriebasierten Nutzermodells einen Gewinn für die Empfehlungsauswahl in Situationen mit spärlichen Datensätzen dar. Die Hypothesen hinsichtlich des Sparsity-Szenarios konnten eindeutig belegt werden. Während der lineare hybride Ansatz im Vergleich zum bewertungsbasierten Ansatz signifikant besser Bewertungen vorhersagen konnte, erzielte der hybride Filter mit Merkmalserweiterung bei der Einschätzung der Relevanz der Objekte signifikant bessere Ergebnisse. Ein völliger Verzicht auf die bewertungsbasierte Ähnlichkeit wie im rein auf die Energiekultur basierendem Verfahren stellte sich allerdings ebenfalls für beide Evaluationsmaße als signifikant schlechter als die hybriden Verfahren heraus.

Auch für das New-User-Szenario konnte gezeigt werden, dass die Berücksichtigung der Energiekulturen der Nutzer von Vorteil sein kann. So verbesserte sich

durch die Filtertechnik mit Merkmalserweiterung die Vorhersage der Bewertungen signifikant, so dass Hypothese 3 bestätigt werden konnte. Für die Korrektheit der Klassifizierung der Objekte nach ihrer Relevanz für die Nutzer konnte lediglich bei Nutzern mit nur einer Bewertung ein deutlicher Unterschied festgestellt werden. Bei zusätzlichen Bewertungen, die in einem realen System höchstwahrscheinlich sehr früh abgegeben werden würden, stellte das Hinzuziehen der Energiekultur hinsichtlich dieses Qualitätskriteriums keinen nennenswerten Vorteil mehr dar, so dass Hypothese 4 durch die Evaluation widerlegt wurde.

4.3.4 Diskussion

Die Evaluationen in den beiden Anwendungsszenarios haben gezeigt, dass durch die Integration anwendungsspezifischer Nutzermodelle, die auf sozialwissenschaftlichen Theorien beruhen, die Qualität der Empfehlungsauswahl in Cold-Start-Szenarien signifikant verbessert werden kann. Dies gilt für die Vorhersage von Bewertungen sowie für die Klassifikation der möglichen Empfehlungen hinsichtlich ihrer Relevanz für die Nutzer. Die besten Ergebnisse konnten durch die hybriden Filterverfahren (lineare Kombination, Merkmalserweiterung) erzielt werden.

Trotz der viel versprechenden Ergebnisse, müssen jedoch einzelne Limitierungen angesprochen werden, die es in den Evaluationen gab und die die Übertragbarkeit der Ergebnisse auf ein reales Empfehlungssystem einschränken.

Da für die untersuchten Anwendungsszenarien bisher keine größeren Datensätze verfügbar sind, wurden die Evaluationen mit einem selbst erstellten und relativ kleinen Datensatz durchgeführt. So kam es, dass der Datensatz im SavER-Szenario zum Beispiel für teure und aufwendige Aktivitäten wie bauliche Maßnahmen oder die Anschaffung eines Hybrid- oder Elektroautos hauptsächlich negative Bewertungen enthielt. Die meisten der Teilnehmer der Online-Umfrage konnten oder wollten sich solch große Investitionen nicht leisten. Realistischerweise muss allerdings davon ausgegangen werden, dass normalerweise der Anteil der Nutzer, für die auch größere Investitionen interessant sind, größer ist. Die Effekte, die solche Gegebenheiten auf die Qualität der Filtertechniken haben, konnten anhand der gesammelten Daten in den Evaluationen nicht vollständig nachvollzogen werden.

Eine weitere Limitierung der durchgeführten Evaluationen ist die fehlende Dynamik in den Nutzermodellen. Eine Annahme bzw. ein Ziel beratender Empfehlungssysteme ist, dass sich die Meinungen und Präferenzen der Nutzer mit der Zeit und im besten Fall durch die Nutzung der Systeme verändern. Die Auswirkungen solcher Anpassungen auf die Performanz der Filterverfahren konnten in den durchgeführten Evaluationen ebenfalls nicht simuliert und untersucht werden.

Im Hinblick auf beide Limitierungen ist jedoch davon auszugehen, dass die Stärken theoriebasierter Nutzermodelle gerade unter den beschriebenen Gegebenheiten zum Tragen kommen sollten.

4.4 Nutzerzentrierte Entwicklung eines CARE-Prototypen

Innerhalb des CARE-Projekts wurde mittels eines nutzerzentrierter Designprozesses ein prototypisches CARE-System entwickelt und evaluiert. Da dieser Prozess interessante Einblicke in die Umsetzung eines beratenden Empfehlungssystems und speziell die Anforderungen der Zielgruppe an die Empfehlungsauswahl gewährte, werden im Folgenden die einzelnen Entwicklungsschritte und die darin gewonnenen Erkenntnisse beschrieben. Im Fokus standen vor allem die Fragen, hinsichtlich welcher Themen sich die Senioren eine Unterstützung durch ein proaktive Empfehlungen vorstellen könnten, wann und wie oft sie mit dem System interagieren würden und wie eine typische Interaktion mit dem System aussehen sollte (z.B. Anzahl und Auswahl der Empfehlungen).

Die in diesem Kapitel beschriebenen Entwicklungsschritte umfassen die Anforderungsanalyse mit einer Nutzerbefragung und die iterative Implementierung des CARE-Systems mit einer zwischenzeitlichen zweiwöchigen Evaluation des ersten Prototypen in der Wohnung eines älteren Pärchens.

4.4.1 Anforderungsanalyse

Grundlegende Anforderungen an die Empfehlungsauswahl förderte bereits die Literaturrecherche, u.a. mit den Modellen aus Kapitel 2.1, zu Tage.

Potential Jede ausgesprochene Empfehlung muss einen erkennbaren Beitrag zur Steigerung des Wohlbefindens leisten können.

Relevanz Empfohlene Maßnahmen müssen in der gegebenen Situation für die Nutzer sinnvoll sein. Zum Beispiel können bestimmte kontextuelle Gegebenheiten (z.B. schlechte Luft) gezielte Handlungen (z.B. Lüften) erfordern, während andere Kontexte (z.B. schlechtes Wetter) bestimmte Handlungen (z.B. Spaziergänge) automatisch ausschließen.

Ausführbarkeit Jede empfohlene Maßnahme muss möglichst sofort durchzuführen sein. Die Voraussetzungen hierfür können sowohl die Nutzer als auch den situativen Kontext (z.B. benötigte Gegenstände) betreffen.

Motivation Es muss absehbar sein, dass die Zielperson in der gegebenen Situation dazu motiviert ist, die empfohlenen Maßnahmen durchzuführen.

Für eine Anforderungsanalyse wurden 27 Senioren (15 Frauen, zwölf Männer) im Alter von 59 bis 92 Jahren befragt. Ihnen wurden kurze comichafte Konzeptvideos, siehe Abbildung 4.20, gezeigt, die verschiedene Szenarien mit einem CARE-System darstellten. Anschließend wurden die Teilnehmer hinsichtlich ihrer generellen Bedürfnisse und Einstellungen gegenüber solchen Systemen interviewt.

Ein Großteil der Senioren schätzte die Idee eines Systems zur Unterstützung im Alltag positiv oder zumindest neutral ein. Danach gefragt, welche Funktionen ihnen

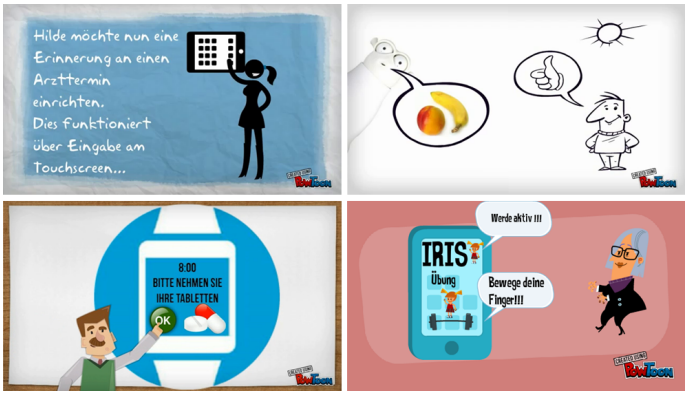

Abbildung 4.20: Screenshots aus den Konzeptvideos, die den Teilnehmern der Anforderungsanalyse gezeigt wurden.

wichtig wären, nannten die meisten Teilnehmer Erinnerungsfunktionen für Termine, Medikamente oder das Ausschalten von Haushaltsgeräten (v.a. sicherheitskritische Geräte wie Herde). Aber auch Empfehlungen zur körperlichen Ertüchtigung und für eine gesunde Ernährung wurden häufig genannt. Um eine bessere Anpassung der Empfehlungen zu ermöglichen, zeigten sich die Senioren auch durchaus offen gegenüber der Erfassung von Kontextinformationen. Allerdings wurde die Aufzeichnung von Kameradaten grundsätzlich abgelehnt, da sie für die Senioren einen zu starken Einschnitt in ihre Privatsphäre darstellte. Viele der Senioren äußersten außerdem Bedenken, dass sie vom System bevormundet oder abhängig werden könnten. Dies unterstrich die Wichtigkeit die Entwicklung des Systems vor allem auf die Steigerung des Selbstbewusstseins und der Motivation der Nutzer auszurichten und diese Ziele den Senioren gegenüber auch klar zu kommunizieren.

Die Eindrücke, die in der Anforderungsanalyse gewonnen werden konnten, wurden durch eine von den griechischen Projektpartnern im CARE-Projekt durchgeführte Befragung auch für griechische Senioren bestätigt [Hammer et al., 2015b].

4.4.2 Implementierung und Evaluation eines ersten Prototypen

Das Ziel des ersten Prototypen war es erste Eindrücke darüber sammeln, wie gut sich das System in den Alltag der Senioren integrieren lässt, wie eine typische Interaktion der Nutzer mit dem System aussieht bzw. aussehen sollte und ob ein System wie CARE überhaupt akzeptiert werden würde.

Da für diese Untersuchungen nur eine rudimentäre Empfehlungsauswahl von Nöten war, wurde lediglich ein einfacher regelbasierter Ansatz genutzt. Dieser sollte grundsätzlich unangebrachte Empfehlungen wie Spaziergänge bei schlechtem Wetter ausfiltern und einzelne Maßnahmen und Aktivitäten wie Lüften bei schlechter Luft oder Entspannungs- und Lachübungen bei schlechter Stimmung gezielt auswählen. Ein ähnlicher Ansatz zur Integration von Empfehlungen für einfache körperliche Aktivitäten in den Alltag der Nutzer wurde von Lin und Kollegen [Lin et al., 2011] mit guten Ergebnissen eingesetzt.

Im ersten Prototypen wurden die folgenden Kontextinformationen verwendet:

- Tageszeit: *Morgen, Vormittag, Mittag, Nachmittag, Abend, Nacht*
- Zeiten für Sonnenaufgang und -untergang
- Wetter: *sehr gut, gut, neutral, schlecht, sehr schlecht*
- Wohnraumklima (basierend auf Raumtemperatur, Luftfeuchtigkeit und Luftqualität): *sehr gut, gut, neutral, schlecht, sehr schlecht*
- Stimmung der Nutzer (basierend auf regelmäßigen Befragungen in Abständen von minimal einer Stunde jeweils zu Beginn der Nutzung des Systems, siehe Abbildung 4.21): *gut, neutral, schlecht*

Abbildung 4.21: Screenshot der Abfrage der persönlichen Stimmung

Die Architektur des ersten CARE-Prototypen sowie die verwendeten Sensoren und die Interpretation der Sensorrohdaten sind in [Rist et al., 2015, Seiderer et al., 2015] detailliert beschrieben. An dieser Stelle liegt der Fokus rein auf der Empfehlungsauswahl und der Evaluation der präsentierten Empfehlungen.

Regelbasiertes Filterverfahren Die zur Umsetzung komplexer Regelsysteme verwendeten *Rule Engines* sind häufig als Open-Source und für verschiedenste Programmiersprachen verfügbar. Eine Übersicht über Open-Source-Rule-Engines für Java-Projekte findet sich unter Java-Source.net[17]. Ein prominentes Beispiel ist die JRuleEngine[18], die es ermöglicht, Regeln auch in XML-Dateien zu verfassen. In CARE wurde ein bereits vorhandenes, in SWI-Prolog[19] implementiertes Regelsystem eingesetzt, dass mittels JPL[20] integriert werden konnte.

Code-Beispiel 4.1 zeigt die Repräsentation der Empfehlung „Raus ins Grüne“ innerhalb des Regelsystems. Jede Repräsentation einer Empfehlung bestand aus einem Namen, einer ID, einer Auflistung der Kategorien von Wohlbefinden, denen die Empfehlung zugeordnet war, und den Bedingungen für die Auswahl der Empfehlung. Diese Bedingungen setzten sich aus Regeln für die jeweils relevanten Kontextinformationen („facts“) zusammen.

[17] http://java-source.net/open-source/rule-engines
[18] http://jruleengine.sourceforge.net
[19] http://www.swi-prolog.org
[20] http://www.swi-prolog.org/packages/jpl

Code-Beispiel 4.1: Beispiele für die Darstellung der Voraussetzungen für die Empfehlung „Raus ins Grüne“ im Regelsystem

```
name:'Raus ins Grüne',
id:36,
categories:[physicalWB,exercises],
conditions:[
  facts:season in [summer, spring],
  facts:timeOfDay in [morning,forenoon,afternoon],
  facts:outdoorBrightness == bright,
  facts:weatherCondition in [veryGood,good],
  facts:userActivity in [low, neutral]]
```

Aufgrund der vereinfachten Darstellung des Kontextes waren für die Formulierung der Regeln lediglich zwei Operatoren notwendig. Mittels „*in*“ und einer eckigen Klammer wurde eine Menge von Zuständen definiert, von denen lediglich einer zutreffen musste. Mit „==“ wurde dagegen ein einzelner zulässiger Zustand für den entsprechenden Kontext festgelegt. Das Beispiel in Code-Beispiel 4.1 zeigt eine Empfehlung, die lediglich im Frühling und Sommer untertags (außer zur Mittagszeit) ausgesprochen werden darf, wenn es hell ist und die Wetterbedingungen mindestens gut sind. Außerdem sorgt die Einschränkung für den Grad der Aktivität dafür, dass diese Empfehlung nicht ausgesprochen wird, wenn eine Person in der letzten Zeit bereits relativ aktiv war.

Bei der Erstellung der Regeln wurden das Wissen und gewisse Theorien über die einzelnen Kategorien von Wohlbefinden so auf die Bedingungen der Maßnahmen und Aktivitäten übertragen, dass diese möglichst immer zu einem geeigneten Zeitpunkt ausgewählt wurden. Beispielsweise wurden Ernährungstipps zu den typischen Essenszeiten und vormittags, wenn meistens das Essen für den Tag geplant wird, ausgesprochen. Körperliche Aktivitäten wurden zu Zeiten ausgeschlossen, in denen die Nutzer sich entspannen sollten (mittags, abends und nachts). Für die Kategorie „Umgebung“ wurden mehrere Varianten der Empfehlung „Lüften“ erstellt, die lediglich bei einer schlechten Luftqualität präsentiert wurden. Die verschiedenen Varianten zeigten unterschiedliche Bilder und Erklärungen, die je nach Jahres- und Tageszeit relevant waren. Empfehlungen zur Steigerung des sozialen Wohlbefindens waren außer nachts immer möglich, da verschiedenste Arten von Kontaktaufnahmen, von Unternehmungen in der Stadt bis zu Telefonaten, empfohlen werden konnten. Diese brachten jeweils ihre eigenen Ausschlusskriterien mit sich.

Zum Abgleich der aktuellen Kontextinformationen mit den Regeln wurden diese Informationen innerhalb des Regelsystems in einer den Regeln ähnlichen Form repräsentiert, siehe Code-Beispiel 4.2. Ein näherer Blick auf die beschriebene Situation zeigt, dass die Jahres- und Tageszeit sowie die Aktivität der Person für die Empfehlung „Raus ins Grüne“ sprechen würden. Allerdings würde die Empfehlung aufgrund schlechter Wetterbedingungen ausgefiltert.

Code-Beispiel 4.2: Beispiele für die Darstellung der Voraussetzungen für die Empfehlung „Raus ins Grüne“ im Regelsystem

```
facts:[
  season:summer,
  timeOfDay:afternoon
  weatherCondition:bad,
  userActivity:neutral,
  userMood:bad,
  roomTemperature:good,
  roomHumidity:good]
```

Falls in einer Situation keine Empfehlung gezielt ausgewählt werden konnte, wurde nach der Ausfilterung der inadäquaten Aktivitäten zufällig eine der restlichen Aktivitäten ausgewählt. Hierfür gab es lediglich die Einschränkung, dass eine Empfehlung nicht mehr als drei Mal innerhalb einer Stunde erscheinen sollte. Durch diese Maßnahme sollte trotz des rudimentär gehaltenen Empfehlungsauswahl eine ausreichende Diversität erreicht und das System interessant gehalten werden.

Die Entscheidung, nur eine einzige Empfehlung für einen vordefinierten Zeitraum (z.B. körperliche Übungen = 60s, Ernährungstipps = 30s) anzuzeigen, beruhte darauf, dass die Interaktion der Senioren mit dem Bilderrahmen, durch den Verzicht auf eine Navigation zwischen mehreren Empfehlungen, so einfach wie möglich gehalten werden sollte. Es sollte lediglich möglich sein, Quizfragen und Fragen zur eigenen Stimmung zu beantworten und die jeweilige Empfehlung mit „gefällt mir“ oder „gefällt mir nicht“ zu bewerten. Auf diese Weise sollte den Nutzern die Angst genommen werden, dass sie an der Bedienung des Systems scheitern könnten.

Die abgegebenen Bewertungen für die Empfehlungen wurde im ersten Prototypen nicht für die Generierung späterer Empfehlungen verwendet. Sie dienten lediglich der späteren Analyse der Nutzerpräferenzen und der Qualität der ausgesprochenen Empfehlungen. Basierend auf dieser Analyse sollte das Empfehlungssystem für den zweiten Prototypen überarbeitet und erweitert werden.

Ebenfalls nur für die Evaluation wurde die Kamera des Bilderrahmens genutzt, um in Absprache mit den Studienteilnehmern im Empfehlungsmodus einzelne Fotos von den Reaktionen der Nutzer aufzunehmen, siehe Abbildung 4.22. Auf eine weitere Nutzung der Kamera wurde zum Schutz der Privatsphäre der Nutzer verzichtet.

Evaluation des ersten Prototypen Die Evaluation des ersten Prototypen fand im Haushalt einer 76-jährigen Frau und eines 75-jährigen Mannes statt und dauerte zwei Wochen. Zu Beginn erfolgte die Installation des Systems und eine kurze Einweisung. Die Evaluation endete mit dem Abbau des Systems und einer abschließenden Befragung. Während der Testphase beschränkten die Evaluatoren Besuche vor Ort auf zwei Termine zur Überprüfung des Systemstatus und zur Sicherung der gesammelten Log-Dateien.

Als geeigneter Standort für das System stellte sich die Küche heraus, in der sich beide Senioren regelmäßig und gerne aufhielten. Dort wurde der Bilderrahmen so angebracht, dass er sich gut sichtbar auf Augenhöhe befand und nicht von weiteren Möbelstücken, Türen oder Fenstern verdeckt wurde, siehe Abbildung 4.22.

Abbildung 4.22 zeigt ein Beispiel für eine Nutzung des Systems während der Testphase. In diesem Fall erhielt der männliche Studienteilnehmer die Empfehlung, die Fenster zu öffnen, um frische Luft herein zu lassen. Die Reaktion in Abbildung 4.22 (d) sowie die gespeicherte Bewertung des Nutzers zeigten, dass ihm die Empfehlung in dieser Situation nicht gefallen hat.

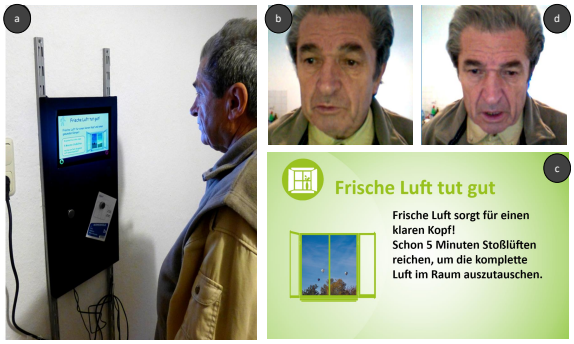

Abbildung 4.22: Interaktion eines Nutzers mit dem CARE-System: Er bleibt vor dem Bilderrahmen stehen (a) und erhält die Empfehlung die Fenster zum Lüften zu öffnen (b,c). Die Reaktion deutet auf eine eher negative Reaktion hin (d).

Resultate Während der Testphase wurden 171 Empfehlungen angezeigt. Die Interaktionen mit dem System verteilten sich über den kompletten Tag, fanden aber vermehrt mittags und nachmittags statt. Von den 49 vorliegenden Empfehlungen wurden 45 mindestens einmal angezeigt, siehe Abbildung 4.23. Das zeigt, dass die Mechanismen zur Steigerung der Vielfalt der Empfehlungen trotz ihrer Einfachheit erfolgreich waren. Die Empfehlung zu lüften wurde mit Abstand am häufigsten (13x) angezeigt. Allerdings führten zwischenzeitliche Probleme mit dem Luftqualitätssensor zu einer häufigeren Auswahl dieser Empfehlung. Ebenfalls häufig wurden Empfehlungen zum Halsdehnen (9x), zwei Gedächtnisübungen (Telefonnummer merken, Heimweg erkennen) und eine Übung zum bewussten Lächeln (je 8x) angezeigt. Die restlichen Empfehlungen wurden relativ gleichmäßig verteilt ausgesprochen.

Leider wurden nur 9% der Empfehlungen von den Senioren bewertet (13x positiv, 2x negativ). Die meisten positiven Bewertungen wurden für körperliche (5) und mentale Übungen (5) abgegeben. Eine mögliche Ursache für die geringe Anzahl an Bewertungen war, dass es immer wieder technische Probleme bei der Eingaben per Touch gab. Die Beschleunigungssensoren des Bilderrahmens registrierten häufiger mutmaßliche Eingaben der Nutzer, die allerdings vom System nicht erkannt wurden. Ob dies durch die allgemein etwas schwergängige Toucheingabe des Tablet PCs und/oder die häufigen Schwierigkeiten von Senioren mit Touchinterfaces [Motti et al., 2013] bedingt war, konnte im Nachhinein nicht geklärt werden.

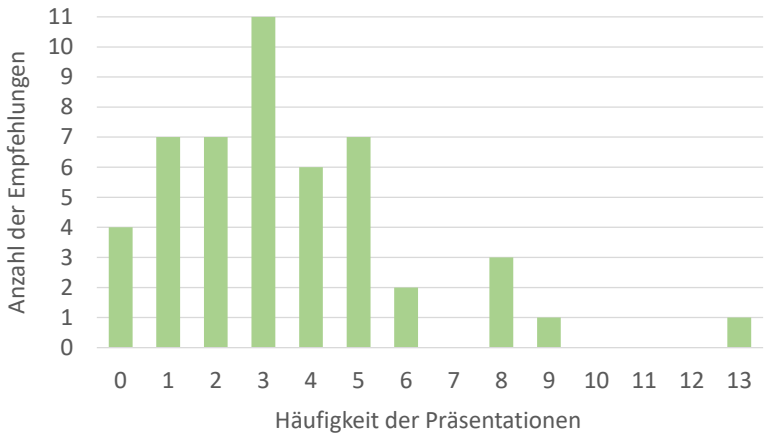

Abbildung 4.23: Häufigkeit der Präsentation einzelner Empfehlungen

In diesem Zusammenhang sei auf eine Studie verwiesen, die im ForGenderCare Projekt durchgeführt wurde, die allerdings kein Bestandteil dieser Dissertation ist [Hammer et al., 2017]. In ihr wurde die Usability und UX von Tablet PCs und sozialen Robotern im CARE-Szenario verglichen. Obwohl die Senioren in dieser Studie nur die Präsentation von Empfehlungen erlebten, wurde der Roboter als signifikant besser nutzbar wahrgenommen als das Tablet PC.

Auf Grund der fehlenden Bewertungen entstammten die wichtigsten Erkenntnisse der Evaluation den abschließenden Interviews. In diesen Interviews wurden anhand von Heuristiken zur Evaluation von Ambient Displays [Mankoff et al., 2003] und Pervasive Health Technologien [Kientz et al., 2010] die subjektiven Eindrücke der Senioren erfragt. Außerdem hatten diese die Möglichkeit zusätzliche Anmerkungen zu äußern. Die Ergebnisse der Befragung werden im Folgenden zusammengefasst.

Integration des Systems in den Alltag der Nutzer (Heuristiken: „Appropriate Time and Place“ und „Peripherality of Display“) Die Stelle, an der das System installiert wurde, wurde von beiden Senioren als gut geeignet empfunden. Obwohl der Bildschirm durchgängig eingeschaltet war, fühlten sie sich nie von seiner Helligkeit gestört. Dies kann u.a. dadurch erklärt werden, dass eine Küche meist nur bei hellen Lichtverhältnissen genutzt wird. In anderen Räumen wie Wohnzimmer oder Schlafzimmer, die auch im Dunklen genutzt werden, könnte das System dagegen als störend wahrgenommen werden [Consolvo und Towle, 2005].

Auch die Präsentation der Empfehlungen konnte durch die Installation in der regelmäßig genutzten Küche meist zu einem geeigneten Zeitpunkt erfolgen. Allerdings hätte die weibliche Teilnehmerin anstatt vieler, über den Tag verteilter, kurzer Interaktionen lieber weniger (z.B. je einmal morgens und abends) und dafür längere Interaktionen durchgeführt. Begründet war dies vor allem durch ihre Präferenzen für körperliche Übungen. Der männliche Teilnehmer, der hauptsächlich Inhalte wie Witze und Scherzfragen bevorzugte, die unabhängig von der Tageszeit angezeigt werden konnten, äußerte keine Vorlieben für die Anzahl und Dauer der Interaktionen.

Interaktion mit dem System (Heuristiken: „Ease of Use“ und „Flexibility and Efficiency of Use“) Hinsichtlich der Darstellung der Empfehlungen waren sich beide Senioren einig, dass die Bilder und Texte gut verständlich und lesbar waren. Auch der Umfang der dargestellten Information wurde als angemessen bewertet. Allerdings muss darauf hingewiesen werden, dass die weibliche Teilnehmerin bereits durch die Teilnahme an einer Reha und durch ihre Tätigkeit als Pflegerin in einem Altenheim ausreichend Vorwissen hatte, um die Auswirkungen der körperlichen Übungen auf ihr Wohlbefinden zu verstehen.

Die Bedeutung der vorhandenen Buttons war den Teilnehmern, laut eigener Angaben, ebenfalls klar. Allerdings deuteten Beobachtungen und die Analyse der gesammelten Sensordaten auf einige Probleme bei der Nutzung des Bilderrahmens hin. Zum Beispiel trat hin und wieder das Problem auf, dass das System zu spät auf die Anwesenheit der Nutzer reagierte. In diesen Situationen gingen die Senioren bereits wieder vom Bilderrahmen weg, als die Inhalte angezeigt wurden. Dies führte teilweise zu Irritationen. Um dieses Missverständnis aufzulösen wäre ein zusätzliches, einfaches Feedback von Seiten des Systems hilfreich gewesen.

Entgegen der Annahme, dass die stark eingeschränkten Interaktionsmöglichkeiten die Akzeptanz des Systems verbessern würden, wünschten sich die Senioren mehr Möglichkeiten das System zu kontrollieren. Sie schlugen u.a. vor, dass das System pausiert oder Aktivitäten auf einen späteren Zeitpunkt verschoben werden können. So könnten für Übungen benötigte Gegenstände wie Bälle oder Stühle geholt werden, falls sie nicht in Reichweite sind, und es könnten Unterbrechungen durch andere Personen, Telefonate oder ähnliches abgefangen werden.

Nützlichkeit und Relevanz der Empfehlungen (Heuristiken: „Useful and relevant information“) Beide Senioren konnten sich hauptsächlich an Empfehlungen für körperliche und emotionale Übungen sowie an Gedächtnis- und Denksportaufgaben erinnern. Dagegen gaben sie an, dass sie sich an keine Empfehlung zum Lüften erinnern könnten. Da diese Empfehlung allerdings, auch aufgrund zwischenzeitlicher technischer Probleme, häufiger angezeigt wurde, könnte es sein, dass sich die Beiden durch die Empfehlungen etwas peinlich berührt fühlten.

Bezüglich der Präferenzen für die einzelnen Kategorien unterschieden sich die beiden Teilnehmer. Während der Mann sich hauptsächlich über Witze und Scherzfragen freute, bevorzugte die Frau körperliche Übungen und dabei im Speziellen Übungen für ihre Finger und Hände. Außerdem war sie sehr von den präsentierten Entspannungsübungen angetan, von denen sie zwar schon von anderen Personen gehört hatte, sie aber bisher nie selbst ausprobiert hatte.

Hinsichtlich der Nützlichkeit der Empfehlungen bemängelten beide Senioren, dass pro Interaktion lediglich eine einzelne Empfehlung angezeigt wurde. Speziell bei körperlichen Übung hätte es, ihrer Meinung nach, mehr Sinn gemacht, mehrere Übungen hintereinander anzuzeigen und so ein kleines Trainingsprogramm von bis zu 20 oder 30 Minuten Länge anzubieten. Außerdem empfanden sie die Anzeigedauer der Empfehlungen generell als zu kurz. Diese Nutzungserfahrung konnte durch

die Log-Daten der Infrarotsensoren bestätigt werden. In mehreren Fällen verharrten die Nutzer länger vor dem Bilderrahmen, als die Empfehlung angezeigt wurde. Dies könnte ein Indiz dafür sein, dass die Zeiten für die Anzeige von Empfehlungen speziell für körperliche und mentale Übungen tatsächlich zu kurz bemessen war und die Teilnehmer selbstständig die Übungen beenden mussten.

Beide Teilnehmer wiesen außerdem darauf hin, dass sie sich insgesamt eine etwas größere Vielfalt an Empfehlungen wünschen würden. Dazu ist allerdings anzumerken, dass zum Zeitpunkt der ersten Evaluation lediglich rund 50 Empfehlungen zur Verfügung standen.

Generelles Fazit Insgesamt äußerten sich beide Senioren überwiegend positiv über das System. Es integrierte sich gut in ihren Alltag und speziell die weibliche Teilnehmerin zeigte sich positiv überrascht. Die körperlichen Übungen und die Entspannungsübungen passten sehr gut zu ihrer starken intrinsischen Motivation, aktiv zu bleiben und Trägheit zu vermeiden. Das System hielt für sie dementsprechend einen größeren Mehrwert bereit als für ihren Mann.

Die Evaluation zeigte aber dennoch, wie erwartet, einiges Verbesserungspotential und nötige Anpassungen an die Nutzerbedürfnisse auf:

1. Mehr Empfehlungen einer Kategorie pro Nutzung
2. Mehr Interaktionsmöglichkeiten zur Navigation zwischen Empfehlungen (z.B. Vor, Zurück, Pause)
3. Robustere Interaktion (u.a. zur Abgabe von Bewertungen)
4. Größere Anreize Empfehlungen anzunehmen (z.B. Anzeige des Fortschritts)
5. Erhöhte Vielfältigkeit der Empfehlungen
6. Erweiterung der Filtertechnologien, um Empfehlungen besser auf die einzelnen Nutzer zuschneiden zu können
7. Personalisierte und situativ angepasste Empfehlungsinhalte (z.B. Argumente oder Erklärungen)

4.4.3 Implementierung des zweiten Prototypen

Die Ergebnisse der Evaluation des ersten Prototypen führten zu Änderungen betreffend der Hardware, des Konzepts des Systems und auch der Empfehlungsauswahl. Abbildung 4.24 zeigt den neugestalteten Prototypen.

Hardware Aufgrund der technischen Probleme mit dem im ersten Prototypen eingesetzten schwächeren Tablet (HP Omni 10 5600eg) wurde für den zweiten Prototypen ein leistungsstärkeres und zuverlässigeres MS Surface III Tablet eingesetzt. Außerdem wurden als Ergänzung zu den Software-Buttons, Hardware-Buttons mit

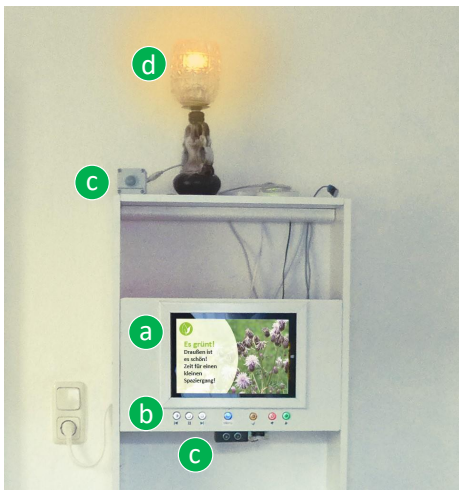

Abbildung 4.24: Überarbeiteter Prototyp des CARE-Systems: (a) Tablet mit Empfehlung, (b) Hardware-Buttons, (c) Sensoren zur Präsenzerkennung, (d) Lampe für „Ambient-Modus“

der selben Funktionalität ergänzt. Die Nutzer hatten damit die Möglichkeit auf zwei Arten mit dem System zu interagieren und somit auch zwei Möglichkeiten eine Bewertung für die aktuelle Empfehlung abzugeben.

Zusätzlich wurden einige Sensoren ergänzt: Zur Stabilisierung der Präsenzerkennung wurden neben dem bisherigen Infrarotsensor auch ein Ultraschallsensor, ein Helligkeitssensor und die in das Tablet eingebauten Mikrophone berücksichtigt. Die Kontextinformationen wurden durch die Messungen eines Kontaktsensors und eines Strommessers (Smartmeter) ergänzt. Zum einen konnte so erkannt werden, ob Fenster geöffnet waren. Zum anderen konnte aufgrund der Betriebszeiten von Geräten wie dem Fernseher auf die Aktivität der Nutzer geschlossen werden. Zu guter Letzt wurde ein Bewegungsmelder zur Erkennung von Personen im gesamten Raum ergänzt. Die dadurch gewonnen Daten wurden für einen neu implementierten Ambient-Modus, siehe Konzeptänderungen, verwendet.

Konzept Um den Nutzern bereits bei Betreten des Raumes signalisieren zu können, dass eine neue Empfehlung vorliegt, wurde ein sog. „Ambient-Modus“ entwickelt. Sobald eine Person den Raum betrat, in dem das System installiert wurde, wurde der Empfehlungsprozess gestartet und der Person über eine zusätzlich installierte Lampe (Philips Hue) signalisiert, dass eine neue Empfehlung für sie vorliegt. Durch eine gezielte Einfärbung der Lampe in der Farbe der Kategorie der ausgewählten Empfehlungen, wurde auch bereits diese Information an die Person weitergeleitet. Dadurch sollte ihre Entscheidung erleichtert werden, ob sie das System in dieser Situation nutzen möchte oder nicht.

Als zusätzliche Kategorie an Inhalten, die allerdings unabhängig vom Empfehlungssystem zu Beginn jeder Nutzung angezeigt wurde, wurden Postkarten ergänzt, die Bilder und kurze Textnachrichten enthielten, die Verwandte und Freunden über eine Smartphone-App erstellen und an das System schicken konnten.

Eine der wichtigsten Konzeptänderungen war, dass pro Nutzung des Systems mehrere Empfehlungen der selben Kategorie angezeigt wurden anstatt wie bisher nur eine einzelne Empfehlung. Zwischen diesen Empfehlungen konnten die Nutzer beliebig vor und zurück navigieren. Eine vordefinierte Anzeigedauer pro Empfehlung gab es nicht mehr. Außerdem wurde ein Menü ergänzt, über das die Senioren auch gezielt Kategorien auswählen konnten. Die Dauer der Nutzung des Systems wurde somit alleine von den Nutzern bestimmt.

Empfehlungsauswahl Als Reaktion auf das geänderte Konzept und den Wunsch der Nutzer nach besser personalisierten und vielseitigeren Empfehlungen, wurde das Empfehlungssystem zu einem kaskadierenden hybriden Empfehlungssystem erweitert, siehe Abbildung 4.25. Nach der bereits beschriebenen kontextbewussten regelbasierten Filterung ungeeigneter Empfehlungen, wurde mit Hilfe einiger weiterer kontextbasierter Regeln eine einzelne Kategorie ausgewählt, aus der Aktivitäten empfohlen werden sollten. Diese Kategorie konnte auch manuell über das neue Menü des Systems ausgewählt werden. In einem letzten Schritt wurden aus den noch übrigen Aktivitäten mittels eines fallbasierten Filters die Aktivitäten ausgewählt, die letztendlich am besten zur aktuellen Situation passten. Die maximale Anzahl an Empfehlungen pro Session wurde auf fünf limitiert, um nicht zu viele Objekte zur Auswahl zu stellen.

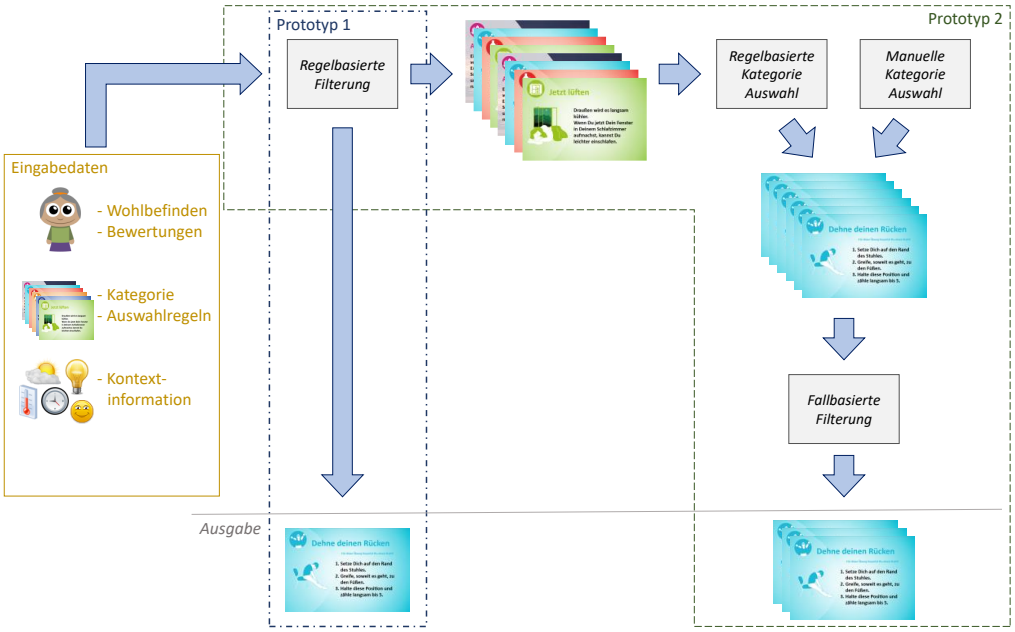

Abbildung 4.25: Wissensbasierte Empfehlungsauswahl in CARE

Regelbasierte Kategorieauswahl Eine Erkenntnis der ersten Evaluation war, dass die Senioren statt einzelner Empfehlungen mehrere Empfehlungen einer Kategorie erhalten wollten. Zu diesem Zweck wurde ein regelbasiertes Filterverfahren implementiert, das die Nützlichkeit der jeweiligen Kategorien im vorliegenden

Kontext einschätzte und die am besten geeignete Kategorie auswählte. Die Regeln für die Auswahl der Kategorien wurden extern in einer json-Datei definiert und waren damit schnell und einfach anpassbar.

In den Regeln wurden für jede Kategorie ausschlaggebende Kontexte mit den Gewichten 1 oder −1 versehen, die die Nützlichkeit der Kategorie in diesem Kontext definierten, siehe zum Beispiel Code-Beispiel 4.3. Körperliche Übungen sollten zum Beispiel vor allem vormittags und nachmittags empfohlen werden und wenn sich die Nutzer bis zu diesem Zeitpunkt zu wenig bewegten. Abends und nachts sollten dagegen keine körperlichen Übungen empfohlen werden, da sich die Nutzer dann entspannen und ausruhen sollten. Empfehlungen zur Förderung des emotionalen Wohlbefindens sollten vermehrt zu späteren Tageszeiten und bei einer schlechten Stimmung der Nutzer präsentiert werden. Nach Abgleich aller relevanten Kontexte mit den Regeln und der Addition aller Gewichte wurde die Kategorie mit der höchsten Nützlichkeit ausgewählt. Bei Gleichstand wurde zufällig entschieden.

Zur Steigerung der Vielseitigkeit der Empfehlungen wurden Kategorien pro Auswahl mit Penaltygewichten von −1 versehen, die bei erneuten Interaktionen innerhalb einer Stunde bei der Bewertung der Nützlichkeit mit eingerechnet wurden.

Code-Beispiel 4.3: Beispiele für Regeln zur Einschätzung der Nützlichkeit einer Kategorie in der gegebenen Situation

```
{
 "name":"exercises",
 "contexts":{
  "timeOfDay:forenoon":1,
  "timeOfDay:afternoon":1,
  "timeOfDay:evening":-1,
  "timeOfDay:night":-1,
  "userActivity:low":1}
}
```

```
{
 "name":"emotionalWB",
 "contexts":{
  "timeOfDay:noon":1,
  "timeOfDay:evening":1,
  "timeOfDay:night":1,

  "userMood:bad":1}
}
```

Fallbasiertes Filterverfahren Die restlichen Aktivitäten der ausgewählten Kategorie wurden in einem letzten Schritt in einem fallbasierten Verfahren hinsichtlich ihrer Nützlichkeit in der aktuellen Situation eingeschätzt.

Schritt 1: Retrieve Zur Berechnung der Ähnlichkeiten zwischen der aktuellen Situation s_{jetzt} und einer früheren Situation s_{alt} wurde für jeden Kontext c mit Gleichung (4.6) die Ähnlichkeit zwischen den Kontextwerten c_{jetzt} und c_{alt} berechnet.

$$sim_c(c_{jetzt}, c_{alt}) = 1 - d(c_{jetzt}, c_{alt})/d_{max} \tag{4.6}$$

c_{jetzt} und c_{alt} erhält man durch die Abbildung der geordneten Kontextwerte auf aufsteigende Zahlenwerte (z.B. „gut, neutral, sehr gut“ auf „1,2,3“). $d(c_{jetzt}, c_{alt})$ wurde mit dem generischen Abstandsmaß in Gleichung (4.7) berechnet.

$$d(c_{jetzt}, c_{alt}) = \begin{cases} |c_{jetzt} - c_{alt}| & \text{falls Kontext nicht zyklisch} \\ & \text{falls Kontext zyklisch} \\ & \text{und } |c_{jetzt} - c_{alt}| <= d_{max} \\ -|c_{jetzt} - c_{alt}| + 2 * d_{max} & \text{falls Kontext zyklisch} \\ & \text{und } |c_{jetzt} - c_{alt}| > d_{max} \end{cases} \tag{4.7}$$

Zyklische Kontexte waren die Tages- und Jahreszeiten. d_{max} berechnete sich sowohl für Gleichung (4.6), als auch für Gleichung (4.7) wie folgt:

$$d_{max} = \begin{cases} c_{max} - c_{min} & \text{falls Kontext nicht zyklisch} \\ c_{max}/2 & \text{falls Kontext zyklisch} \end{cases} \tag{4.8}$$

Die Ähnlichkeit zwischen zwei Situationen berechnete sich dann mittels des gewichteten arithmetischen Mittels:

$$sim_S(s_{jetzt}, s_{alt}) = \frac{\Sigma_{c \in C} sim_c(c_{jetzt}, c_{alt}) * w_c}{\Sigma_{c \in C} w_c} \tag{4.9}$$

Da in dieser Entwicklungsphase noch keine Erfahrungswerte hinsichtlich der Wichtigkeit einzelner Kontexte für die Empfehlungsauswahl vorhanden waren, wurden die Gewichte w_c für alle Kontexte auf 1 gesetzt.

Schritt 2: Reuse Die Nützlichkeit $N_{s,i}$ einer Aktivität i in der aktuellen Situation s_{jetzt} berechnete sich anhand der gewichteten Summe aller Bewertungen, die in früheren Situation S für die Aktivität abgegeben wurden.

$$N_{s,i} = (\Sigma_{s \in S} sim_s(s_{jetzt}, s_s) * r_{s,i}) - P(i) \tag{4.10}$$

$P(i)$ war ein Penalty der, analog zur Auswahl der Kategorie, pro Empfehlung von i in der letzten Stunde um 1 erhöhte wurde. Dies sollte wiederum zu abwechslungsreicheren Empfehlungen führen. Die Aktivitäten, die $N_{s,i}$ maximierten, wurden in der aktuellen Situation empfohlen.

Schritt 3: Revise Da die für die Filterung ausschlaggebenden Faktoren zum größten Teil aus auf Sensordaten basierenden Kontextinformationen bestanden, war eine direkte Anpassung der Anforderungen an die Aktivitäten in diesem Fall nicht möglich. Die einzige Möglichkeit, die Empfehlungsauswahl maßgeblich zu manipulieren, wäre die Auswahl einer anderen Kategorie gewesen. Dies war bereits über das neu hinzugefügte Menü möglich.

Schritt 4: Retain Nach jeder Anzeige einer Empfehlung wurde für diese Empfehlung ein Eintrag in der Falldatenbank gespeichert. Dieser Eintrag enthielt die ID der Empfehlung, den aktuellen Kontext und die Bewertung der Nutzer (1 = gut, -1 = schlecht oder 0 = nicht abgegebenen).

4.4.4 Diskussion

Die nutzerzentrierte Entwicklung des CARE-Systems zeigte, dass es durchaus das Potential hat, von Senioren akzeptiert und genutzt zu werden. Die gewonnen Erkenntnisse stammen zwar bisher nur aus einer Evaluation mit zwei Teilnehmern. Diese Evaluation fand allerdings über einen Zeitraum von zwei Wochen und in situ statt, wodurch die Ergebnisse und Einblicke sehr wertvoll sind.

Durch die Evaluation des ersten Prototypen wurde auch deutlich, dass eine nutzerzentrierte Vorgehensweise beim Design und der Entwicklung beratender Empfehlungssysteme unabdingbar ist. Die Konzeptänderungen, die bei der Entwicklung des zweiten Prototypen vorgenommen wurden, zeigten, dass die vorherigen Annahmen der Entwickler, die auf einer Literaturrecherche und Anforderungsanalyse beruhten, in mehreren Punkten nicht mit den tatsächlichen Anforderungen während der Nutzung des CARE-Systems übereinstimmten. Von der Platzierung des Systems im Zuhause der Senioren, über die akzeptierte Sensorik, über die Anzahl der Empfehlungen pro Nutzung oder pro Tag, bis hin zu weiteren erwünschten Funktionalitäten gab es vielfältige Faktoren, die zusammen mit den Nutzern der Zielgruppe untersucht und geklärt werden mussten.

Dass dieses Vorhergehen durchaus positive Folgen auf die Wahrnehmung des Systems hat, zeigte ein erster Test des überarbeiteten Prototypen mit den Teilnehmern der ersten Evaluation. Eine ausgiebige Evaluation des überarbeiteten Prototypen konnte innerhalb dieser Dissertation jedoch nicht mehr durchgeführt werden.

4.5 Zusammenfassung

In diesem Kapitel wurde die Empfehlungsauswahl in assistierenden Empfehlungssystemen in zweierlei Hinsicht erforscht.

Der erste Teil der Untersuchungen befasste sich mit der Verbesserung kollaborativer Filtertechniken durch theoriebasierte Nutzermodelle. Die vielversprechenden Ergebnisse der Untersuchungen zeigten, dass durch die Integration anwendungsspezifischer Nutzermodelle, die auf sozialwissenschaftlichen Theorien beruhen, die Qualität der Empfehlungsauswahl vor allem in Cold-Start-Szenarien signifikant verbessert werden konnte. Dies galt sowohl für die Vorhersage von Bewertungen für Aktionen und Maßnahmen, die den Nutzern zuvor unbekannt waren, als auch für die Klassifikation der möglichen Empfehlungen hinsichtlich ihrer Relevanz für die Nutzer. Die besten Ergebnisse konnten durch hybride Filterverfahren (lineare Kombination, Merkmalserweiterung) erzielt werden.

Die nutzerzentrierte Entwicklung eines CARE-Systems ergab interessante Einblicke in die Anforderungen der Nutzer an ein solches System. Eine wichtige Erkenntnis war, dass die Senioren weniger und dafür länger andauernde Nutzungen mit mehreren Empfehlungen eines Themas hintereinander gegenüber vielen, einzelnen Interaktionen mit einzelnen Empfehlungen bevorzugten. Im Ausblick auf eine längerfristige Motivation zur Nutzung des Systems waren eine große Vielfältigkeit der Empfehlungen und zusätzliche Anreize zur Nutzung und Befolgung der Emp-

fehlungen erwünscht. Neben Fortschrittsanzeigen wurden in diesem Zusammenhang auch personalisierte und situativ angepasste Empfehlungsinhalte wie Argumente oder Erklärungen genannt.

Wie die Personalisierung von Empfehlungstexten aussehen könnte und welche Folgen sie für die Wahrnehmung eines beratenden Empfehlungssystems und seiner Empfehlungen hat, ist Bestandteil des folgenden Kapitels.

5 Generierung von Empfehlungstexten

Menschen verhalten sich gegenüber Computersystemen ähnlich sozial, wie sie es auch gegenüber Menschen tun [Reeves und Nass, 1998]. Daher ist anzunehmen, dass nicht nur durch die Qualität der Empfehlungen, sondern auch durch andere Faktoren die UX mit beratenden Empfehlungssystemen beeinflusst werden kann.

Ein Beispiel sind natürlichsprachliche Empfehlungstexte in mündlicher oder textueller Form. Diese Art der Ausgabe macht für assistierende Empfehlungssysteme zum einen deswegen Sinn, da sie der natürlichen Kommunikationsform der Menschen entspricht und so eine soziale Bindung zwischen Nutzer und System fördern könnte. Zum anderen bietet sich die Möglichkeit u.a. durch die Auswahl überzeugender Erklärungen und Argumente, die Strukturierung des Textes sowie durch die Ausformulierung des Textes die Wirkung der Empfehlungen auf die Nutzer zu beeinflussen [Marcu, 1996]. Allerdings muss beachtet werden, dass natürlichsprachliche Aussagen nicht nur vertrauenswürdig oder überzeugend, sondern zum Beispiel auch einschüchternd oder beleidigend wirken können [Searle, 1969].

In diesem Kapitel liegt der Fokus darauf, wie Empfehlungstexte aufgebaut sein sollten, um das Nutzervertrauen, die Nutzerakzeptanz und die Überzeugungskraft einer Empfehlung zu steigern. Es werden sowohl die Effekte personalisierter Argumente als auch die Wirkung personalisierter und situativer Formulierung von Empfehlungstexten untersucht.

Kulturbasierte Auswahl überzeugender Argumente Durch Erklärungen kann erreicht werden, dass die Nutzer die Hintergründe von Empfehlung besser verstehen und diese anschließend auch annehmen. Ein wichtiger Bestandteil einer Erklärung ist die Argumentation für die Befolgung der jeweiligen Empfehlung. Da Argumente von Person zu Person als unterschiedlich überzeugend wahrgenommen werden, wird in dieser Dissertation untersucht, ob die Nutzer eines assistierenden Empfehlungssystems besser von der Nützlichkeit einer empfohlenen Maßnahme überzeugt werden können, wenn die präsentierten Argumente personalisiert werden. Ein wichtiges Kriterium für die Wahrnehmung von Argumenten ist der kulturelle Hintergrund der Menschen [Aaker und Maheswaran, 1997, Han und Shavitt, 1994, Williams et al., 2006]. Da es bereits fundierte Theorien und Modelle über die Eigenschaften und Werte verschiedener Kulturen gibt [Hofstede, 2001, Hofstede et al., 2010, Triandis, 1995], liegt es nahe, diese Theorien und Modelle für die Auswahl von Argumenten in beratenden Empfehlungssystemen zu nutzen. Ob dies tatsächlich möglich ist und welchen Effekt die kulturbasierte Auswahl von Argumenten auf die Überzeugungskraft von Empfehlungstexten hat, ist Bestandteil der Untersuchungen in diesem Kapitel.

Formulierungen basierend auf soziologischen Theorien Anders als bei der emotional eher neutralen Interaktion mit E-Commerce-Empfehlungssystemen besteht in beratenden Empfehlungssystemen die Gefahr einer emotionalen Barriere

zwischen Nutzer und System. In ihnen weisen die Empfehlungen die Nutzer häufig auf Schwächen oder falsches Verhalten hin. Dadurch könnten Gefühle wie Bevormundung oder Scham ausgelöst werden, die dem Nutzervertrauen und der Nutzerakzeptanz schaden und zu einer Ablehnung des Systems führen können. Eine angepasste Formulierung könnte dies vermeiden.

Welche Wirkung verschiedene Formulierungen der selben Empfehlung haben können, soll am folgenden Beispiel verdeutlicht werden: (1) „Schalte das Licht im Schlafzimmer aus, wenn du es nicht mehr brauchst. Es verbraucht unnötig Energie." (2) „Wolltest du nicht etwas Energie sparen? Du könntest das Licht im Schlafzimmer ausschalten, falls du es nicht mehr brauchst."

Beide Varianten verfolgen dasselbe Ziel und nutzen das selbe Argument. Variante 1 ist recht direkt formuliert und könnte dadurch als extrovertiert, einschüchternd oder bevormundend aufgefasst werden. Variante 2 versucht dagegen, eine direkte Konfrontation zu vermeiden. Sie weist indirekt auf das vorhandene Problem hin und überlässt der Zielperson die Entscheidung. Dadurch könnte sie als höflicher und respektvoller wahrgenommen werden. Allerdings besteht das Risiko, dass diese Variante als weniger überzeugend wahrgenommen wird.

Mairesse [Mairesse, 2008] zählt die Förmlichkeit, die Höflichkeit und die Persönlichkeit, die durch eine Äußerung übertragen wird, sowie verwendete Dialekte und Soziolekte zu den linguistischen Faktoren, die die Wahrnehmung einer Äußerung steuern können. Für die in dieser Arbeit untersuchten Herausforderungen scheinen vor allem die Höflichkeit und die durch Empfehlungstexte vermittelte Persönlichkeit ausschlaggebend zu sein.

Der Grad der Höflichkeit spiegelt wider, welches Maß an Respekt und Selbstwertgefühl einem Konversationspartner entgegengebracht bzw. gewährt wird. Dadurch stellt die gezielte Anwendung von Höflichkeitsstrategien ein probates Mittel dar, um die Chancen zu verbessern, dass die eigenen Kommunikationsziele wie der Aufbau von Vertrauen oder die Überzeugung von einer Sache erreicht werden [Brown und Levinson, 1987]. Inwiefern bekannte Höflichkeitsstrategien auf die Formulierungen in System wie CARE oder SavER übertragbar sind, ist eine Forschungsfrage dieser Dissertation.

Die wahrgenommene Persönlichkeit eines Systems kann ebenfalls auf die Akzeptanz und das Vertrauen der Nutzer einwirken [Reeves und Nass, 1998]. Weist eine Person zum Beispiel eine ähnliche Persönlichkeit wie man selbst auf, wirkt sie häufig vertrauter und die Chancen auf gegenseitige Akzeptanz und gegenseitiges Vertrauen stehen besser. Ob selbiges auch für das Verhältnis zwischen Nutzern und beratenden Empfehlungssystemen gilt und ob die wahrgenommene Persönlichkeit eines Systems eine Auswirkung auf die Überzeugungskraft von Empfehlungen hat, wird ebenfalls in diesem Kapitel untersucht.

Aufbau des Kapitels In Kapitel 5.1 wird zunächst ein Grundverständnis für Erklärungen und Argumente in Empfehlungssystemen vermittelt. Anschließend wird untersucht, ob es einen Effekt auf die Überzeugungskraft von Argumenten hat, wenn

diese basierend auf dem kulturellen Hintergrund der Nutzer ausgewählt werden. In den darauf folgenden Kapiteln wird der Frage nachgegangen, ob durch eine situativ bewusst gewählte Formulierung von Empfehlungstexten gezielte Wirkungen wie das Hervorheben der Wichtigkeit einer Empfehlung oder die Pflege der Beziehung zwischen Nutzer und System erreicht werden kann. Hierfür wird die Wahrnehmung verschiedener Höflichkeitsstrategien, siehe Kapitel 5.2, und verschiedener Ausprägungen der Persönlichkeit des Systems, siehe Kapitel 5.3, verglichen. In allen drei Untersuchungen werden psychologische und soziologische Modelle und Frameworks eingesetzt, die gut in ein Empfehlungssystem integriert werden können und die aufgrund ihrer Theorien über die zwischenmenschliche Kommunikation auf vielversprechende Ergebnisse hoffen lassen.

5.1 Personalisierte Auswahl von Argumenten

Eine Erklärung stellt Zusammenhänge zwischen einzelnen Fakten her, um gewisse Ziele zu erreichen [Friedrich und Zanker, 2011, Tintarev und Masthoff, 2011]. In Empfehlungssystemen beschreiben Erklärungen meist die Zusammenhänge zwischen den Eigenschaften der Empfehlungen und den Präferenzen der Nutzer [Felfernig et al., 2008, Pu und Chen, 2006, Zanker, 2012] oder zwischen den Präferenzen der Zielperson und anderer Nutzer [Bilgic, 2005, Herlocker et al., 2000]. Dadurch können sie maßgeblich zur Akzeptanz eines Empfehlungssystems beitragen [Herlocker et al., 2000].

Ziele von Erklärungen Mit Erklärungen von Empfehlungen können, laut Tintarev und Masthoff [Tintarev und Masthoff, 2011], die folgenden Ziele verfolgt werden:

Transparenz, Verständnis, Validität Tintarev und Masthoff unterscheiden drei Ziele, die gemeinsam haben, dass die Nutzer verstehen sollen, warum sie eine bestimmte Empfehlung erhalten haben.

Die *Transparenz* beinhaltet eine detaillierte Erklärung der Funktionalität des Empfehlungssystems. Durch diese Erklärung sollen die Nutzer nachvollziehen können, warum einzelne Objekte anderen vorgezogen wurden. Oft ist es allerdings nicht möglich oder sogar nachteilig, genau zu erklären, wie das Empfehlungssystem die Empfehlungen ausgewählt hat [Herlocker et al., 2000]. Das *Verständnis* für die Entscheidungen eines Systems kann bereits dadurch gefördert werden, dass man versucht, das mentale Modell der Nutzer („Wie stellen sich die Nutzer die Funktionsweise des Empfehlungssystems vor?“) mit dem Implementierungsmodell des Systems („Wie funktioniert das Empfehlungssystem tatsächlich?“) in Verbindung zu bringen. Ein Beispiel in einem kollaborativen Empfehlungssystem wäre: „Das System hat die Nutzer bestimmt, deren Präferenzen ähnlich zu den Ihren sind. Diese haben die empfohlene Maßnahme als hilfreich bewertet.“ Verwenden Empfehlungssysteme komplexere Algorithmen, ist meist keine einfach verständliche Erklärung der Verfahren mehr möglich. Außerdem kann es sein, dass Nutzer nicht daran interessiert sind,

zu verstehen, wie das Empfehlungssystem tatsächlich funktioniert. In diesen Fällen sollten zum Beispiel Metainformationen wie Kosten oder interessante Kontextinformationen wie das Wetter angezeigt werden, anhand derer die Nutzer abschätzen können, warum die jeweilige Empfehlung gerade für sie unter Berücksichtigung ihrer Präferenzen bzw. ihrer aktuellen Situation nützlich ist (*Validität*).

In vielen Forschungsarbeiten, siehe zum Beispiel [Gedikli et al., 2014, Pu und Chen, 2006, Zanker, 2012], wird die Transparenz als ein allgemeinerer Faktor verstanden, der jedes der drei Ziele von Tintarev und Masthoff umfassen kann. Auch in dieser Arbeit wird der Begriff „Transparenz" so verstanden, dass die Nutzer allgemein die Gründe für das Systemverhalten verstehen. Wie genau das „Warum" erklärt wird, hängt jedoch vom jeweiligen Anwendungsfall und den Nutzern ab und bleibt den Entwicklern zukünftiger Systeme überlassen.

Weiterbildung Erklärt man Nutzern die Zusammenhänge zwischen einzelnen Empfehlungen und dem zugrundeliegenden Nutzermodell und Kontext, so verschafft dies den Nutzern zusätzliches Wissen in der Domäne des Empfehlungssystems. Dadurch können Nutzer zum einen ihre Präferenzen überdenken und gegebenenfalls anpassen und zum anderen ihre Präferenzen gegenüber dem System besser formulieren und so den Empfehlungsprozess stärker beeinflussen.

Effektivität, Effizienz, Zufriedenheit Durch eine gute Erklärung, warum ein bestimmtes Objekt oder eine bestimmte Aktion für sie nützlich ist, können Nutzer einfacher und schneller (*Effizienz*) eine korrekte Einschätzung darüber treffen, welche Empfehlungen für sie relevant sind und welche nicht (*Effektivität*). Es sollte allerdings darauf geachtet werden, dass nicht zu viele zusätzliche Informationen angezeigt werden. Dies könnte zu einem neuerlichen Informationsüberfluss und damit einem höheren mentalen Aufwand der Nutzern führen. Beides sollte durch den Nutzen von Empfehlungssystemen eigentlich verhindert werden. Gelingt es die Effizienz und/oder Effektivität eines Systems durch Erklärungen zu steigern, so führt dies sehr wahrscheinlich auch zu einer größeren *Zufriedenheit* der Nutzer.

Vertrauen, Überzeugungskraft Reduzieren Erklärungen die Ungewissheit über die Qualität der Empfehlungen oder unterstützt man mit ihnen Nutzer erfolgreich bei der Entscheidungsfindung, kann dies das *Nutzervertrauen* gegenüber dem System und seinen Empfehlungen steigern. Kennzeichen für Nutzervertrauen sind Loyalität und im Falle von E-Commerce-Systemen gesteigerte Verkaufszahlen [Tintarev und Masthoff, 2011]. Hinsichtlich des Vertrauens ist interessant, dass vorangegangene Arbeiten zeigten, dass die Art der Objekte bzw. die möglichen Folgen einer schlechten Empfehlung einen Einfluss darauf haben, wie viele Gedanken sich die Nutzer über die Vertrauenswürdigkeit des Systems machen. Während bei einer Studie mit Filmempfehlungen die Art der Erklärung das Vertrauen der Nutzer nur wenig beeinflusste und die Vertrauenswürdigkeit der Empfehlungen insgesamt relativ niedrig bewertet wurde [Tintarev und Masthoff, 2008], zeigten Pu und Chen

[Pu und Chen, 2006], dass sich Nutzer bei teureren Produkten wie Notebooks oder Digitalkameras mehr Gedanken über die Vertrauenswürdigkeit eines Systems machen und dies in Befragungen auch zu aussagekräftigeren Bewertungen hinsichtlich des Nutzervertrauens führte. Die Empfehlungen in CARE und SavER sind zwar meistens nicht mit einem hohen finanziellen Risiko verbunden, aber dafür mit einem höheren körperlichen oder zeitlichen Aufwand. Außerdem kann es die Nutzer Überwindung kosten, Maßnahmen und Handlungen durchzuführen, die für sie ungewohnt sind. Deswegen kann auch für diese Empfehlungssysteme davon ausgegangen werden, dass Nutzer das Vertrauen in das entsprechende System höher gewichten.

Durch Erklärungen kann außerdem die *Überzeugungskraft* eines Empfehlungssystems verbessert werden. Wie stark die Überzeugungskraft einer Erklärung war, kann zum Beispiel dadurch abgeschätzt werden, wie sehr sich die Einschätzung einer Person hinsichtlich eines Objekts durch die Erklärung verbessert hat [Bilgic, 2005]. Da Erklärungen eine Empfehlung auch (absichtlich) besser erscheinen lassen können, als sie tatsächlich ist, wird in Bezug auf das Ziel der Überzeugung allerdings häufig auf ethische Bedenken hingewiesen, siehe z.B. [Ehrlich et al., 2011]. Umso wichtiger war es für diese Dissertation, dass sowohl die Überzeugungskraft als auch das Vertrauen der Erklärungen untersucht und berücksichtigt wurden.

Argumente in Erklärungen Eine entscheidende Rolle in Erklärungen nehmen die präsentierten Argumente für oder gegen die Annahme einer Empfehlung ein. Wie überzeugend ein Argument ist, hängt von seiner Diskrepanz zu den Werten und Meinungen der Nutzer, von seiner Stärke und von der Involvierung der Nutzer in die Argumentation bzw. von der Personalisierung des Arguments ab [Nguyen et al., 2007].

Diskrepanz eines Arguments Die Diskrepanz eines Arguments beschreibt, laut Sherif und Kollegen [Sherif et al., 1981], die relative Distanz zwischen der themenspezifischen Einstellung einer Person und der durch das Argument kommunizierten Meinung. Die maximale Distanz, bis zu der diese Meinung akzeptiert bzw. toleriert wird, ist die *Latitude of Acceptance*. Je näher die tatsächliche Distanz der Latitude of Acceptance kommt, desto wahrscheinlicher ist eine Überzeugung bzw. eine Meinungs- oder Verhaltensänderung. Je stärker ein Verhalten oder eine Meinung in einer Person verankert sind, umso geringer ist jedoch die Toleranzgrenze für andere Meinungen. Ist ein Argument allerdings stark genug, kann es auch dann noch erfolgreich sein, wenn es sich außerhalb der Latitude of Acceptance befindet.

Stärke eines Arguments Die Stärke eines Arguments wird als der Grad definiert, zu dem eine Person das Argument als überzeugend wahrnimmt. Sie hängt von mehreren Faktoren ab. Lee und See [Lee und See, 2004] versuchten zum Beispiel, durch drei Arten von zielorientierten Informationen das Vertrauen der Nutzern zu gewinnen: Informationen über vergangene und aktuelle Ergebnisse, die die Expertise des Systems belegen, Informationen über das generelle Vorgehen des Systems und Informationen über die Ziele des Systems. Gkika und Lekakos

[Gkika und Lekakos, 2014] zeigten, dass soziale Argumente, die die Einschätzungen von Experten oder einer größeren Menge anderer (ähnlicher) Personen hervorheben, stark genug sind, um die Meinung von Nutzern zu ändern. Ein weiterer interessanter Ansatz zur Auswahl von Argumenten ist das sog. *Issue Framing*, bei dem gewisse Faktoren eines Themas stärker hervorgehoben werden, um die Entscheidungsfindung einer Person zu beeinflussen [Wood, 2000]. Im Fokus stehen dabei die Konsequenzen, die ein Verhalten oder eine Entscheidung mit sich bringen können. Während das *Gewinn-Framing* mögliche Vorteile eines Verhaltens unterstreicht, beschreibt das *Verlust-Framing* mögliche Nachteile. Laut Rothman und Salovey [Rothman und Salovey, 1997] sind im Gesundheitswesen zum Beispiel verlustorientierte Nachrichten effektiver, wenn Kontrolluntersuchungen gefördert werden sollen. Gewinnorientierte Nachrichten sind dagegen effektiver, wenn es um präventive Maßnahmen wie das Einhalten von Diäten geht.

Personalisierung eines Arguments Wie die Definitionen der Diskrepanz und der Stärke von Argumenten zeigten, hängt die Überzeugungskraft eines Arguments stark von den individuellen Meinungen, Werten, Zielen und Bedürfnissen der jeweiligen Zielperson ab [Johnson und Eagly, 1989]. Dass auch der kulturelle Hintergrund von Personen die Wahrnehmung von Argumenten und Überzeugungsstrategien beeinflussen kann, zeigen Beispiele aus der Werbung [Aaker und Maheswaran, 1997, Han und Shavitt, 1994] und Erfahrungen aus kulturspezifischen Gesundheitsprogrammen [Williams et al., 2006]. Ein einfaches Beispiel für die Personalisierung von Argumenten in Empfehlungssystemen ist der Bezug auf die Präferenzen der Nutzer oder auf aktuelle, für die Nutzer relevante Kontextinformationen [Bader et al., 2011a, Baltrunas et al., 2011].

Auswahl von Argumenten Für die Auswahl von Argumenten gibt es verschiedene Varianten. Carenini und Moore [Carenini und Moore, 2006] bewerteten zum Beispiel die Nützlichkeit einzelner Argumente und kombinierten die nützlichsten anschließend zu einer Erklärung. Felfernig und Kollegen [Felfernig et al., 2008] bewerteten dagegen nur die Nützlichkeit der kompletten Erklärung. Bader und Kollegen [Bader et al., 2011a] nutzten einen situativen Ansatz. Sie bewerteten sowohl die situative Nützlichkeit einzelner Argumente und Informationen als auch den Informationsgehalt der kompletten Erklärung. Bei der Generierung einer Erklärung ergänzten sie solange nützliche Argumente und Informationen, bis die Nützlichkeit der gesamten Erklärung ausreichend hoch war.

Informationsüberflutung durch Argumente Eine Herausforderung bei der Zusammenstellung von Erklärungen ist, dass durch die präsentierten Argumente keine Informationsüberflutung entstehen darf. Es gilt eine Balance zu finden, die die Nutzer ausreichend informiert, sie aber nicht überfordert. Sollen Erklärungen auf mobilen Interaktionsgeräten mit kleinen Bildschirmen dargestellt werden, erschwert sich diese Aufgabe weiter.

Basierend auf den Ergebnissen verwandter Arbeiten erscheint eine Anzahl von zwei kurzen Argumenten in einer Erklärung als geeignet [Bader et al., 2011a, Herlocker et al., 2000]. Damit Nutzer dennoch bei Bedarf Zugang zu weiteren Informationen bekommen können, bietet sich eine sog. *Ramping Strategy* [Rhodes, 2000] an. Bei dieser Strategie werden mehrere aufeinander aufbauende Detailgrade für Erklärungen verwendet. In der untersten bzw. ersten Stufe werden nur die Informationen direkt mit der Empfehlung angezeigt, mit deren Hilfe das Interesse der Nutzer geweckt und eine schnelle Entscheidungsfindung über die Relevanz der jeweiligen Empfehlung gefördert werden kann. Besteht Interesse an zusätzlichen und detaillierten Erklärungen und Daten, so können die Nutzer bei Bedarf in einem weiteren Schritt weitere Information anfordern.

Forschungsfrage In beratenden Empfehlungssystemen werden Nutzern häufig Aktionen und Maßnahmen empfohlen, die einen zusätzlichen Aufwand oder sogar die Überwindung alter Gewohnheiten erfordern. Deshalb ist es wichtig, die Empfehlungen durch Argumente zu ergänzen, die so überzeugend sind, dass die Nutzer sich überwinden und die empfohlenen Maßnahmen zum Zweck der Förderung ihres Wohlbefindens durchführen. Durch die Analyse verwandter Arbeiten steht bereits fest, dass Erklärungen mit personalisierten Argumenten überzeugender sind als standardisierte Erklärungen. Außerdem führen kulturelle Unterschiede zu unterschiedlichen Wahrnehmungen von Argumenten. Demzufolge sollen in diesem Kapitel die folgenden Fragen beantwortet werden: Können die kulturell bedingten Eigenschaften und Werte der Nutzer mittels Kulturmodellen bei der Auswahl personalisierter Argumente berücksichtigt werden? Hat die kulturelle Auswahl von Argumenten einen Einfluss auf die Überzeugungskraft von Empfehlungstexten?

5.1.1 Kulturmodelle

Die *Kultur* eines Menschen beeinflusst, meist unbewusst, die Wahrnehmung und Auswahl von Verhaltensweisen und Werten. Mit den Energiekulturen in Kapitel 4.2.2 wurde bereits ein auf domänenspezifischen Werten basiertes Kulturmodell vorgestellt. Das klassische Verständnis einer Kultur bezieht sich dagegen auf ethnische Werte und Eigenschaften [Hofstede et al., 2010]. Diese Art der Kultur ist auf spezifische Gruppen wie zum Beispiel die eigene Familie, das soziale Umfeld (z.B. Nachbarschaft, Arbeitsplatz) oder das Heimatland beschränkt. Die im Laufe des Lebens in der jeweiligen Kultur erlernten Verhaltensweisen und Werte können sich zwischen unterschiedlichen Kulturen deutlich unterscheiden.

Mittels Hofstedes *Kulturdimensionen* [Hofstede, 2001, Hofstede et al., 2010] und Triandis *kulturellen Syndromen* [Triandis, 1995] können die Eigenschaften und Werte von Kulturen sowie die Unterschiede zwischen Kulturen beschrieben werden. Die Kulturdimensionen werden zum Beispiel durch Ausprägungen auf Skalen von 0 bis 100 ausgedrückt, so dass unterschiedliche Kulturen einfach beschrieben und miteinander verglichen werden können [Hofstede, 2001].

Individualismus (IND) und Kollektivismus (KOL) Diese Kulturdimension ist die wichtigste, da sie global für die größten Unterschiede zwischen Kulturen verantwortlich ist [Hofstede et al., 2010, Triandis, 1995]. Aus diesem Grund wird sie auch am häufigsten in interkulturellen Untersuchungen verwendet.

Ist in Kulturen individualistisches Denken („Ich“) stärker ausgeprägt, so bestehen zwischen den Menschen nur lose Verbindungen, und es wird davon ausgegangen, dass man sich hauptsächlich um sich und seine unmittelbaren Familienmitglieder kümmert. Individuelle Interessen werden als wichtiger eingeschätzt, als die Interessen der Gemeinschaft bzw. des Kollektivs. Mitglieder individualistischer Kulturen handeln selbstmotiviert, zielorientiert und nach ihrer individuellen Einstellung. Sie beziehen große Motivation aus Schuldgefühlen und dem Verlust von Selbstrespekt, aber auch aus möglichen Vorteilen für sich selbst.

Ist kollektivistisches Denken („Wir“) weiter verbreitet, so herrscht zwischen allen Menschen von Geburt an ein starkes Gefühl der Verbundenheit, und es wird erwartet, dass man sich innerhalb der Verwandtschaft oder anderen Gruppierungen uneingeschränkt loyal verhält und füreinander sorgt. Die Interessen und Bedürfnisse der Gruppe überwiegen gegenüber den individuellen Interessen. Jeder versucht, die Harmonie innerhalb der Gruppe aufrechtzuerhalten und teilt deswegen die traditionellen Fähigkeiten, Eigenschaften und Tugenden der Anderen. Eine Trennung von der Gruppe löst Besorgnis aus. Motivation beziehen die Mitglieder einer kollektivistischen Kultur aus der Schande und dem möglichen Gesichtsverlust, die drohen, falls nicht das getan wird, was innerhalb des Kollektivs als richtig angesehen wird.

Machtdistanz bzw. horizontale und vertikale Beziehungen (MACHT) Eine weitere wichtige Kulturdimension ist die von Hofstede [Hofstede, 2001, Hofstede et al., 2010] definierte *Machtdistanz*. Triandis [Triandis, 1995] spricht in seiner Arbeit von *horizontalen und vertikalen Beziehungen*. Beide meinen allerdings dasselbe. In Gesellschaften mit einem hohen Maß an Machtdistanz bzw. mit vertikalen Beziehungen akzeptieren die Menschen eine hierarchische Ordnung und ihren Platz in dieser Ordnung ohne größere Rechtfertigung. Die Macht ist ungleichmäßig verteilt und Entscheidung werden von „oben“ vorgegeben und nicht diskutiert. Dagegen erstreben Gesellschaften mit einem niedrigen Maß an Machtdistanz bzw. horizontalen Beziehungen eine gleichmäßige Verteilung der Macht. Die Menschen sind weniger abhängig von den „Oberen“ wie z.B. Vorgesetzten und dürfen auch Kritik äußern. Im Falle ungleich verteilter Macht fordern die Menschen Rechtfertigungen für diesen Missstand.

Sowohl Triandis als auch Hofstede haben weitere Kulturdimensionen bzw. kulturelle Syndrome definiert, die sich zum Großteil in beiden Kulturmodellen wiederfinden. Während Triandis allerdings von einer unbekannten Anzahl noch nicht erforschter Syndrome ausging, beschränkte sich Hofstedes Kulturmodell auf lediglich wenige zusätzliche Dimensionen.

Femininität und Maskulinität In maskulinen Kulturen herrscht die klassische Rollenverteilung. Männer haben bestimmt, hart und materiell orientiert zu sein. Frauen sollen dagegen bescheiden und sensibel sein und eine hohe Lebensqualität anstreben. Konkurrenzkampf, Durchsetzungsvermögen, Errungenschaften im Allgemeinen und materielle Belohnungen für Erfolge werden hoch angesehen. In femininen Gesellschaften überschneiden sich die Rollen von Mann und Frau. Männer übernehmen Aufgaben im Haushalt. Frauen dürfen und sollen beruflich erfolgreich sein. Durch stabile Beziehungen, Zusammenhalt und Kooperation, egal ob zuhause oder im Beruf, soll für alle eine hohe Lebensqualität erreicht werden. Angesehene Eigenschaften sind u.a. Bescheidenheit und die Unterstützung Schwächerer.

Vermeidung von Unsicherheit Wie unwohl sich Menschen einer Kultur im Hinblick auf Unsicherheit und Ambiguität fühlen, wird in der Dimension *Vermeidung von Unsicherheit* ausgedrückt. Kulturen mit einem hohen Drang, Unsicherheit zu reduzieren, versuchen das Auftretens unbekannter Situationen zu vermeiden und sind deswegen wenig tolerant gegenüber Veränderungen. Menschen dieser Kulturen scheuen das Risiko zu scheitern. Es wird versucht, unbekannte Situationen vorhersehbar zu machen und sie durch Gesetze, Regeln und strenge Verhaltenskodices zu regulieren. In Kulturen mit einer geringeren Neigung zur Vermeidung von Unsicherheit machen sich die Menschen bedeutend weniger Sorgen um ihre Zukunft. Sie akzeptieren sowohl Unsicherheiten als auch abweichende Meinungen. Veränderungen werden häufig als Chancen für Verbesserungen gesehen. Handlungen zählen, auch wenn sie scheitern, mehr als das Beharren auf Prinzipien.

Lang- (LANG) und Kurzzeitorientierung (KURZ) Diese Dimension, die auf den Lehren und Prinzipien des Konfuzius basiert, befasst sich damit, wie stark sich eine Gesellschaft auf Traditionen und Normen stützt und wie sie mit aktuellen und zukünftigen Herausforderungen umgeht. Langfristig orientierte Kulturen versuchen, Traditionen an neue Gegebenheiten anzupassen und sie für langfristige Ziele zu nutzen. Traditionelle Werte wie harte Arbeit, Loyalität, Beharrlichkeit, Sparsamkeit, aber auch Investitionen werden genutzt, um für die Zukunft eine gute Ausgangslage zu erreichen. Kurzfristig orientierte Kulturen sind dagegen stark durch Tradition und Normen reguliert. Einem gesellschaftlichen Wandel wird in diesen Kulturen skeptisch gegenüber gestanden. Schnelle Resultate und Selbstverwirklichung stehen im Fokus. Hierfür werden Flexibilität, Kreativität und Egoismus als hilfreich angesehen. Der Respekt gegenüber Traditionen, das Erfüllen sozialer Verpflichtungen und die Wahrung des eigenen Ansehens sind von großer Wichtigkeit.

Nachgiebigkeit und Beherrschung Die letzte Dimension befasst sich damit, wie stark es in einer Kultur möglich ist, seinen eigenen grundlegenden, natürlichen Bedürfnissen nachzugehen. Dazu zählen Spaß und Lebensfreude in der Freizeit, das Ausleben der Sexualität oder auch einfache Dinge wie farbenfrohe Kleidung. In Kulturen, in denen diese Möglichkeit relativ frei besteht, bezeichnet sich ein relativ hoher

Prozentsatz der Menschen als optimistisch, glücklich und selbstbestimmt. Zurückhaltende und stark beschränkte Kulturen versuchen dagegen, menschliche Bedürfnisse durch strenge soziale Normen zu kontrollieren. Sparsamkeit, Fleiß und Recht und Ordnung haben eine höhere Priorität als Freizeit und Selbstverwirklichung. Der Blick in die Zukunft ist eher pessimistisch.

In Abbildung 5.1 ist ein beispielhafter Vergleich der deutschen, ägyptischen und US-amerikanischen Kultur anhand Hofstedes Erkenntnissen dargestellt [Hofstede, 2017, Schwartz, 2004]. Für Ägypten sind u.a. eine hierarchische Machtordnung und kollektivistisches Denken charakteristisch. Die eigene Arbeit wird als Mittel gesehen, mit dem das Ziel eines erfüllten Lebens und Wohlbefinden erreicht werden soll. Das Leben ist durch rigide Gesetze und Normen reguliert. Die deutsche Kultur ist eher individualistisch und durch eine eher geringe Machtdistanz geprägt. Das Leben in Deutschland wird durch viele und detailreiche Gesetze und Regeln in geordnete Bahnen gelenkt. Außerdem werden traditionelle Werte wie Fleiß, ein Leben im Einklang mit der Natur, das Streben nach Weltfrieden, Wissensdurst und Kreativität hoch angesehen, da sie dabei helfen, zukünftige Herausforderungen zu meistern und den Wohlstand zu bewahren. Deutsche leben, um zu arbeiten und ziehen ihre Selbstvertrauen aus ihren Aufgaben. Die US-Amerikaner zeichnen sich durch einen Individualismus aus, der stärker ausgeprägt ist als in den meisten anderen westlichen Kulturen. Sie sind eher an kurzfristigen Erfolgen und dem Ausleben ihrer Bedürfnisse interessiert. Sie arbeiten allerdings auch hart dafür. In der US-amerikanischen Kultur ist außerdem ein stark traditionalistisches Denken verbreitet und neue Informationen und Erkenntnisse werden zunächst skeptisch betrachtet und auf ihre Richtigkeit überprüft. Damit im Widerspruch steht ihre „Can-Do"-Einstellung, mit der US-Amerikaner ihre Zukunft anpacken.

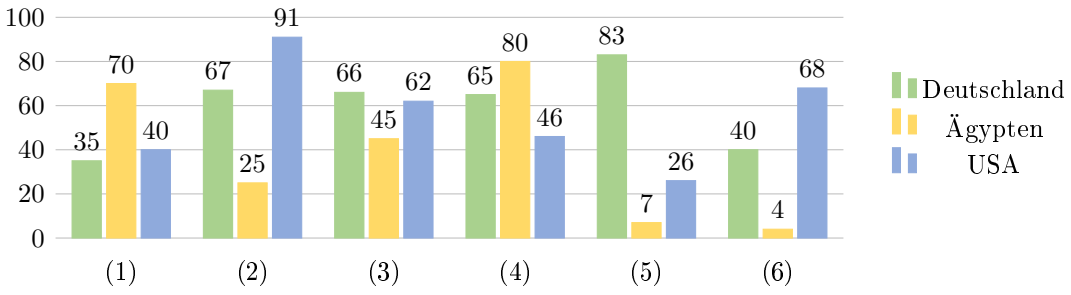

Abbildung 5.1: Vergleich der Kulturdimensionen der Länder Deutschland, Ägypten und USA. (1) Machtdistanz, (2) Individualismus, (3) Maskulinität, (4) Vermeidung von Unsicherheit, (5) Langzeitorientierung, (6) Nachgiebigkeit (nach Hofstede [Hofstede, 2017])

5.1.2 Kulturbasierte Argumente im Anwendungsszenario SavER

Da virtuelle Agenten, die die Kultur der Nutzer reflektieren, von diesen bevorzugt werden [Endrass et al., 2013], erscheint es vielversprechend, die Ausprägungen von

Hofstedes Kulturdimensionen [Hofstede, 2001, Hofstede et al., 2010] in verschiedenen Kulturen auch bei der Auswahl überzeugender Argumente zu berücksichtigen. Khaled und Kollegen [Khaled et al., 2006] entwickelten zum Beispiel Überzeugungsstrategien, die auf den Eigenschaften und Werten kollektivistischer Kulturen beruhten. Dazu gehörten u.a. das Anzeigen von Meinungen anderer Mitglieder der Kulturgruppe, der Bezug positiver und negativer Auswirkungen auf die komplette Gruppe und der Vergleich des eigenen Verhaltens mit dem Verhalten anderer Mitglieder der Kultur, um auf Abweichungen von der Gruppennorm hinzuweisen.

Für gängige Argumente für Energiespartipps können ähnliche Strategien definiert werden. In individualistischen und maskulinen Kulturen sollten Argumente überzeugend sein, die auf ein gesteigertes Ansehen und das Erreichen persönlicher Ziele hinweisen. Sparsamkeit (Geld oder Energie) wird dagegen langfristig orientierten und beherrschten Kulturen zugeschrieben. In Kulturen mit langfristiger Orientierung könnte außerdem mit dem Erhalt der Umwelt argumentiert werden. Bei individualistischen und kollektivistischen Kulturen bietet es sich an, Argumente zusätzlich durch eine Fokussierung auf einen bestimmten Personenkreis zu variieren. Individualistisch denkende Nutzer könnten durch eine Argumentation, die sich auf sie selbst und ihre engere Familie beziehen, überzeugt werden. Bei kollektivistisch denkenden Nutzern sollten dagegen Argumente, die sich auf einen erweiterten Freundes- und Bekanntenkreis oder allgemein die Gemeinschaft beziehen, eine stärkere Wirkung haben. Beispielhafte Argumente für Energiesparaktionen sind in Tabelle 5.1 aufgelistet.

Tabelle 5.1: Argumente für Maßnahmen zum Energiesparen und ihre kulturelle Ausprägung bzgl. der Kulturdimension Individualismus (IND)/Kollektivismus (KOL)

Thema	Beispielargument
Geld/IND	Dadurch könntest Du für Dich und Deine Familie Geld ansparen.
Geld/KOL	Dadurch könntest Du für Deine Freunde und Verwandten Geld ansparen.
Ansehen/IND	Deine Familie würde Dich als umweltbewusst wahrnehmen.
Ansehen/KOL	Freunde und Verwandte würden Dich als umweltbewusst wahrnehmen.
Umwelt/IND	Damit würdest Du Deinen Beitrag dazu leisten, die Umwelt für Deine Kinder und Enkel zu bewahren.
Umwelt/KOL	Damit würdest Du Deinen Beitrag dazu leisten, die Umwelt für kommende Generationen zu bewahren.
Ziele/IND	Damit würdest Du Deine persönliche Umweltbilanz verbessern.
Ziele/KOL	Damit würdest Du Deinen Beitrag dazu leisten, dass die nationalen Emissionsziele erreicht werden können.

5.1.3 Online-Studie im Anwendungsszenario SavER

Um die Annahme zu bestätigen, dass Hofstedes Kulturmodell eine gute Grundlage darstellt, um kulturabhängig überzeugende Argumente für Energiespartipps auszuwählen, wurde eine Online-Studie durchgeführt. Da das Ziel der Studie eine Befragung einer größeren Anzahl an Menschen in verschiedenen Regionen der Welt war und lediglich die Einschätzung der Studienteilnehmer hinsichtlich verschiedener Argumente untersucht werden sollte, wurde auf eine technisch aufwendige, prototypische Umsetzung eines SavER-Systems verzichtet.

Hypothesen

1. Menschen verschiedener Kulturen halten unterschiedliche Argumente für die Durchführung von Energiesparaktionen für überzeugend.

2. Die Argumente, die in einer spezifischen Kultur als überzeugend wahrgenommen werden, spiegeln die in Hofstedes Kulturmodell beschriebenen Eigenschaften und Werte der jeweiligen Kultur wider.

Studienablauf Die Online-Umfrage wurde in Englisch durchgeführt und enthielt neben demographischen Fragen bzgl. des Alters, der Nationalität und des Geschlechts der Befragten zwei weitere Fragebögen:

1. Verfassen eigener Argumente In diesem Fragebogen bestand die Aufgabe der Teilnehmer darin, für acht vorgegebene Empfehlungen, siehe Tabelle 5.2, überzeugende Argumente für Mitglieder der eigenen Kultur zu formulieren. Um eine möglichst variantenreiche Auswahl von Empfehlungen (einfach/aufwendig und üblich/unüblich) zu erhalten, wurden je zwei Empfehlungen aus den in Kapitel 4.3.3 identifizierten Gruppen von Energiesparaktionen gewählt. Außerdem wurden aus den Bereichen Heizenergie, Stromverbrauch und Benzinverbrauch in etwa gleich viele Empfehlungen ausgewählt. Bei der Auswahl wurde darauf geachtet, dass Empfehlungen des selben Bereichs nie aus der selben Gruppe stammten. Das Ziel dieses Teils der Studie war es herauszufinden, welche Argumente die Studienteilnehmer ohne eine Beeinflussung durch Beispiele auswählen würden.

2. Bewertung vorgegebener Argumente Im zweiten Teil der Befragung wurden die selben Empfehlungen nochmals präsentiert, dieses Mal allerdings jeweils mit den acht verschiedenen Argumenten aus Tabelle 5.1 (in Englisch).

Diese Kombinationen aus Empfehlungen und Argumenten sollten die Teilnehmer auf einer Likert-Skala von 1 = „not persuasive“ (nicht überzeugend) bis 5 = „strongly persuasive“ (sehr überzeugend) bewerten. Außerdem gab es die Möglichkeit, die abgegebenen Bewertungen zu begründen. Die Annahme war, dass Teilnehmer unterschiedlicher Kulturen, im Falle der Gültigkeit der Hypothesen, die Argumente als unterschiedlich überzeugend wahrnehmen würden.

Tabelle 5.2: Liste der in der Online-Umfrage angezeigten Empfehlungen

Gruppe	Bereich	Empfehlung
(1)	Heizung	Turn down the heating in unused rooms...
(1)	Strom	Light up only the rooms that are occupied...
(2)	Benzin	Go by bike or foot more often...
(2)	Strom	Replace your old, normal bulbs with energy-saving lamps...
(3)	Heizung	Buy programmable thermostats for your heaters...
(3)	Benzin	Use public transport more often...
(4)	Benzin	Buy an electric or hybrid car...
(4)	Heizung	Install heat-reflective mats behind radiators...

Studienteilnehmer Die Online-Studie wurde per E-Mail, über soziale Netzwerke und in Foren zum Thema „Energiesparen“ international verteilt. Es nahmen 51 Männer und 43 Frauen aus 15 Ländern an der Studie teil, siehe Abbildung 5.2. Zirka die Hälfte der Teilnehmer war jünger als 30. Etwa 30% der Teilnehmer waren mittleren Alters. Die restlichen Teilnehmer waren 50 oder älter.

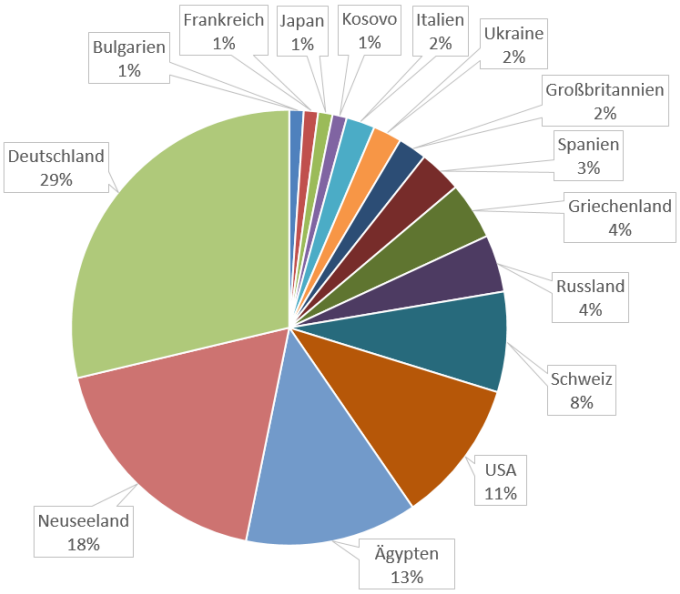

Abbildung 5.2: Überblick der Nationalitäten der Studienteilnehmer

Ergebnisse: Verfassen eigener Argumente Insgesamt wurden 1009 Argumente für die Annahme der acht Empfehlungen genannt. Sie wurden nach ihrem Thema (z.B. Geld, Energie, Umwelt) und, falls vorhanden, nach ihrer expliziten Ausprägung hinsichtlich einer Kulturdimension (z.B. Langzeitorientierung, Kollektivismus) kategorisiert. Abbildung 5.3 zeigt die Verteilung der Argumente auf die Kategorien.

Häufig wurden Argumente naheliegender Themen wie *Geld-* oder *Energiesparen* sowie der *Schutz und Erhalt der Umwelt* genutzt. Oft wurde auch angedeutet, dass die genannte Maßnahme eigentlich *logisch* sein sollte. Außerdem wurden bei den

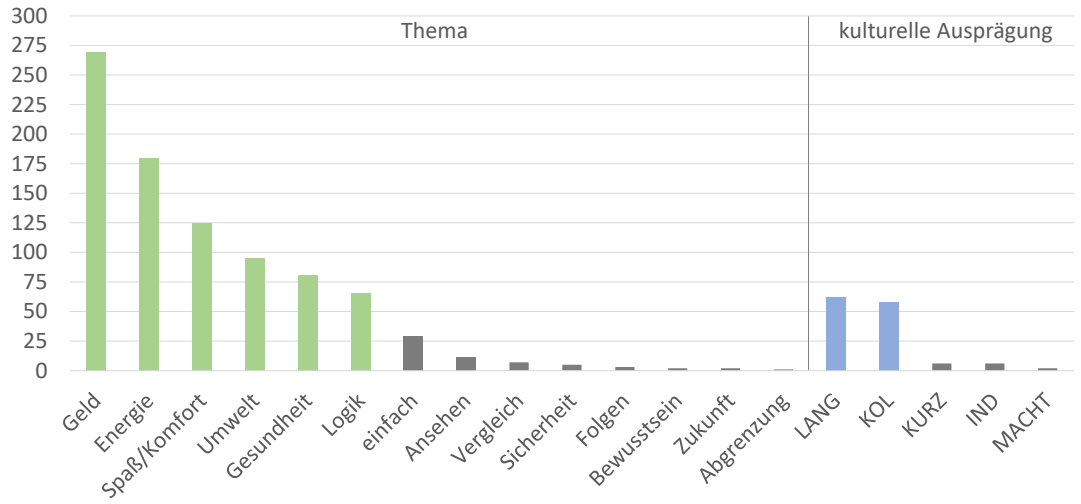

Abbildung 5.3: Auflistung und Anzahl aller genannten Argumente (grau: nicht für die statistische Auswertung berücksichtigt)

Empfehlungen zum Benzinsparen häufig themenspezifische Argumente aus den Kategorien *Gesundheit* und *Spaß bzw. Komfort* angesprochen. Beispiele für Argumente der genannten Themen sind in Tabelle 5.3 aufgelistet.

Tabelle 5.3: Beispiele häufig verwendeter Argumente

Kategorisierung nach Themen	
Geld	„save money“, „it’s cheaper“, „save/reduce costs“
Energie	„save energy“, „use energy more efficiently“
Spaß/Komfort	„reduces traffic in rush hours“, „more time to relax“, „feel comfortable“, „a lot of fun, driving a electrocar“
Umwelt	„better for the environment“, „reduce co2-emission“, „it’s environment-friendly“
Gesundheit	„could prevent mold“, „it’s healthier“, „do sports“,
Logik	„it doesn’t make sense to light up/heat unoccupied rooms“, „it’s common sense“
Kategorisierung nach der expliziten kulturellen Ausprägung	
LANG	„new lights last longer“, „it’s cheaper on the long run“, „if more people buy them, they will get affordable for the majority“
KOL	„(help to) save the planet“, „save the resources of our country“, „we have lots of power outages because...“

Die nur vereinzelt genutzten Themen umfassten u.a. die *einfache Umsetzung* einzelner Energiespartipps, das *Ansehen*, das durch die Anschaffung eines Hybrid- oder Elektroautos erreicht werden könnte, und der *Vergleich* mit Anderen. Auffällig war, dass beinahe alle Argumente nur positive Auswirkungen erwähnten.

Einige der Argumente konnten direkt auf die Kulturdimensionen *Individualismus* (z.B. „it saves your own money“, „you will be different from others“), *Kollektivismus* (siehe Tabelle 5.3), *Langzeitorientierung* (siehe Tabelle 5.3), *Kurzzeitorientierung* (z.B. „you will notice significant differences within one month“) und *Machtdistanz* (z.B. „should be mandatory“, „our president recommends using bikes“) abgebildet werden. Allerdings kamen diese Argumente nicht zwingend von Teilnehmern aus entsprechend orientierten Kulturen. Allgemein deuteten die ersten Ergebnisse darauf hin, dass die Teilnehmer unabhängig von ihrer Kultur ähnlich argumentierten.

Vergleich individualistisch und kollektivistisch orientierter Kulturen Für die tiefergehende statistische Auswertung wurden nur Kategorien herangezogen, die häufiger als 50-mal genannt wurden ($> 5\%$ der Gesamtmenge an Argumenten), siehe Abbildung 5.3 (grün und blau). Da nur Deutschland, Neuseeland, Ägypten und die USA mit 10 oder mehr Teilnehmern vertreten waren, wurden die Herkunftsländer der Teilnehmer wie in vielen kulturellen Studien üblich für die Analyse zunächst nach der Kulturdimensionen Individualismus/Kollektivismus gruppiert, siehe Tabelle 5.4. Da es für den Kosovo bisher noch keine Einschätzung durch Hofstede gibt, wurde diese einzelne Person in der weiteren Analyse nicht berücksichtigt.

Um herauszufinden, ob es kulturelle Unterschiede bei der Auswahl der Argumente gab, wurde verglichen, wie groß die Wahrscheinlichkeit war, dass eine Person aus einer der beiden Kulturgruppen ein Argument aus einer der Kategorien nutzen würde. Hierfür wurde für die beiden Gruppen pro Argumentkategorie der Anteil der Teilnehmer bestimmt, die ein Argument der Kategorie genutzt hatten. Für die Überprüfung auf signifikante Unterschiede zwischen den Kulturgruppen wurden für alle Argumentkategorien Chi-Quadrat-Tests durchgeführt.

Tabelle 5.4: Einordnung der Herkunftsländer der Studienteilnehmer nach Individualismus (IND) und Kollektivismus (KOL)

Gruppe	Länder (Anzahl der Teilnehmer aus dem jeweiligen Land)
KOL	Ägypten (12), Bulgarien (1), Griechenland (4), Japan (1), Russland (4), Ukraine (2)
IND	Deutschland (27), Frankreich (1), Großbritannien (2), Italien (2), Neuseeland (17), Schweiz (7), Spanien (3), USA (10)

Die Ergebnisse dieser Tests bestätigten die bisherigen Eindrücke. Für keine der untersuchten Argumentkategorien konnte eine signifikante Verbindung zwischen der kulturellen Einordnung nach IND/KOL und der Wahrscheinlichkeit für die Nutzung von Argumenten aus einer der Kategorien festgestellt werden.

Vergleich der Kulturen hinsichtlich der Einordnung nach allen Hofstede-Dimensionen Nach den Ergebnissen des ersten Vergleichs wurden die Chi-Quadrat-Tests mit einer neuen Gruppierung der Länder anhand aller Hofstede-Dimensionen wiederholt. Ein kMeans-Clustering ergab die in Tabelle 5.5 dargestellte

vier Gruppierung der Länder. Da C2 nur einen Teilnehmer enthielt, wurde dieses Cluster in der statistischen Analyse nicht berücksichtigt. Für die drei verbliebenen Kulturgruppen ergab sich eine signifikante Verbindung bezüglich der Verwendung langfristig orientierter Argumente ($\chi^2(1) = 7,233, p < 0.05$). Bei Teilnehmern aus Ländern der Cluster C1 und C3 war die Chance auf die Wahl eines langfristig orientierten Arguments mehr als drei Mal (C1-C4: 3,75; C3-C4: 3,44) so groß, als bei Teilnehmern aus C4.

Tabelle 5.5: Clustering der Herkunftsländer der Studienteilnehmer nach allen Kulturdimensionen von Hofstede); in Klammern: Anzahl der Teilnehmer

Zuteilung	Länder
C1	Ägypten (12), Bulgarien (1), Griechenland (4), Russland (4), Spanien (3), Ukraine (2)
(C2)	(Japan (1))
C3	Großbritannien (2), Neuseeland (17), Schweiz (7), USA (10)
C4	Deutschland (27), Frankreich (1), Italien (2)

Diskussion Die Analyse der selbstständig verfassten Argumente zeigte kaum kulturelle Unterschiede. Über alle Kulturen hinweg wurden häufig naheliegende Argumente wie Geld- oder Energiesparen oder der Erhalt der Umwelt genannt. Für einzelne Empfehlungen wurden teilweise auch themenspezifische Argumente genutzt. Beispiele waren die Förderung der Gesundheit beim Fahrradfahren oder zu Fuß gehen und ein komfortableres Pendeln mit öffentlichen Verkehrsmitteln.

Eine signifikante Verbindung zwischen dem kulturellem Hintergrund und der Wahrscheinlichkeit der Nutzung von Argumenten einer bestimmten Kategorie lieferte ein Vergleich, in dem die Herkunftsländer der Teilnehmer hinsichtlich aller Kulturdimensionen gruppiert wurden, siehe Tabelle 5.5. In den Clustern C1 und C3 war die Chance, dass ein Argument mit einer langfristigen Orientierung gewählt wurde, signifikant größer als im Cluster C4. Vergleicht man die Länder in den Gruppierungen allerdings hinsichtlich ihrer Lang- bzw. Kurzzeitorientierung, siehe Abbildung 5.4, sind diese Ergebnisse zunächst nicht direkt zu erklären. Alle Länder in C4 sind nach Hofstede als langfristig orientiert einzuschätzen, während von den Ländern in C3 lediglich die Schweiz eine langfristige Orientierung aufweist. Die meisten Teilnehmer in C1 sind ebenfalls eher kurzfristig orientiert.

Etwas Aufklärung verschafft ein Blick auf die Themen der 65 Argumente, die auf langfristige Effekte hinwiesen. Beinahe die Hälfte (46%) der Argumente bezog sich auf finanzielle Ersparnisse. Auffallend ist dabei, dass in C1 43% und in C3 45% aller langfristig orientierten Argumente auf eine langfristige Geldersparnis hinwiesen, während es für die Länder in C4 nur 19% waren. Womöglich war es Mitgliedern der kurzfristig orientierten Länder in C1 und C3 aufgrund des kurzfristigen finanziellen Aufwands mancher der empfohlenen Maßnahmen ein spezielles Bedürfnis auf die langfristig finanziellen Vorteile hinzuweisen.

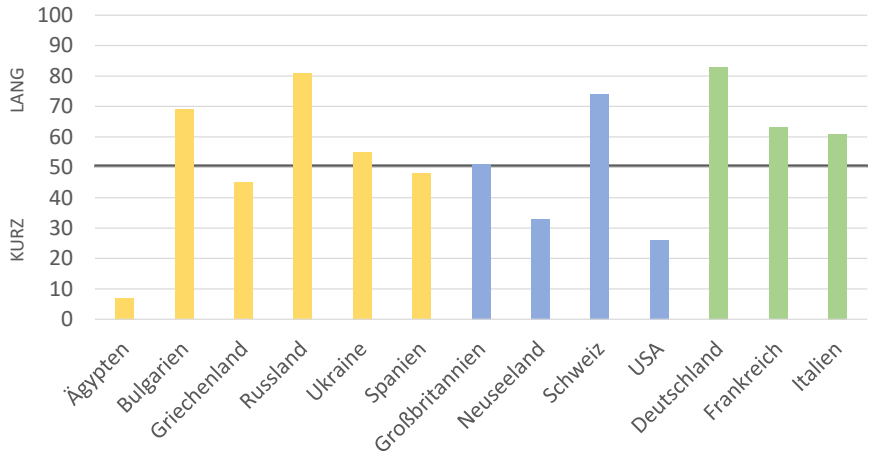

Abbildung 5.4: Hofstedes Bewertung der Länder hinsichtlich der Kulturdimension Lang- und Kurzzeitorientierung (gelb = Cluster C1, blau = C3, grün = C4)

Ähnlich verhielt es sich mit der langfristigen Auswirkung der Maßnahmen auf den Komfort der Nutzer, dem Thema, das am zweithäufigsten gewählt wurde. Etwa ein Drittel (34%) der Argumente aus C3 wiesen darauf hin, dass trotz eines kurzfristigen Aufwands (z.B. Anschaffung und Installation energiesparender Lampen) auf lange Sicht Vorteile (z.B. Geldersparnisse) entstehen könnten. Teilnehmer aus C4 nutzen dieses Argument immerhin in einem Viertel der Fälle. In C1 betrug der Anteil 17%.

Interessanterweise bildeten Argumente, die langfristige Vorteile für die Umwelt, die Mitmenschen oder zukünftige Generationen betreffen, in den Clustern C1 und C3 jeweils nur noch einen Anteil von 13%. In den laut Hofstede eigentlich auf die Zukunft fokussierten Ländern in C4 betrug ihr Anteil dagegen 50%.

Insgesamt konnte die Hypothese, dass Menschen verschiedener Kulturen unterschiedliche Argumente für Energiespartipps für überzeugend halten, nur teilweise belegt werden. Die Hauptbeweggründe für die Durchführung von Energiesparmaßnahmen waren interkulturell dieselben (u.a. Geld- und Energiesparen). Die Unterschiede zwischen den Kulturen lagen in den Feinheiten und können womöglich nicht oder zumindest nicht allein durch Hofstedes Kulturdimensionen begründet werden.

Ergebnisse: Bewertung vorgegebener Argumente Zur Auswertung dieses Teils der Studie wurden alle Teilnehmer wieder nach Individualismus und Kollektivismus gruppiert, siehe Tabelle 5.4. Anschließend wurden in den beiden Kulturgruppen die durchschnittlichen Bewertungen für die Argumente der Themen Strom, Heizung und Benzin ermittelt.

Abbildung 5.5 deutet bereits an, dass sich die Ergebnisse des ersten Teils der Studie bestätigen könnten. Es ist kein eindeutiger Unterschied zwischen individualistischen und kollektivistischen Kulturen zu erkennen. In beiden Gruppen hatte das Argument, dass man für sich selbst Geld sparen könnte, die größte Überzeugungskraft. Am schlechtesten schnitten das Erreichen individueller Zielsetzungen sowie die generelle Verbesserung des Ansehens ab.

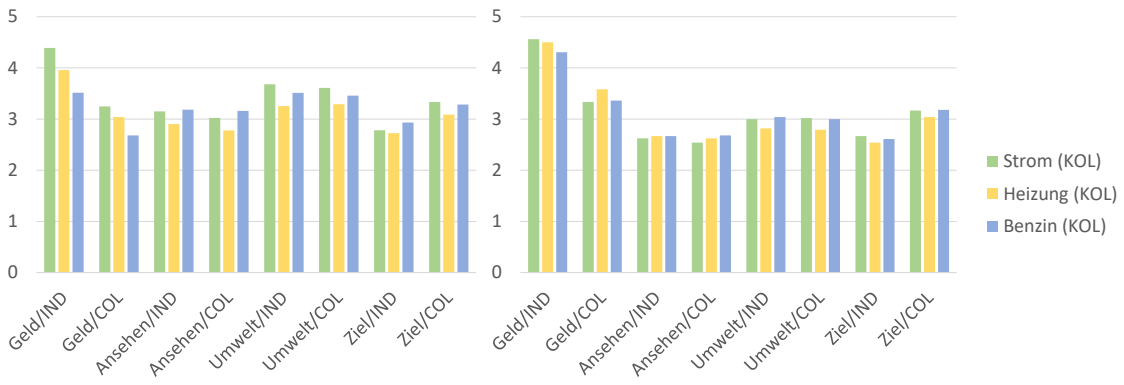

Abbildung 5.5: Durchschnittlich wahrgenommene Überzeugungskraft der Argumente pro Empfehlungsthema (Strom, Heizung und Benzin). Links: individualistische Kulturen, Rechts: kollektivistische Kulturen

Für die statistische Auswertung wurde ein Mixed ANOVA-Test mit einem Bonferroni-Post-Hoc-Test durchgeführt. Der Mixed ANOVA-Test war nötig, da vier unabhängige Variablen die wahrgenommene Überzeugungskraft beeinflussen hätten können: Thema der Empfehlung (Heizung, Strom, Benzin), Thema des Arguments (Geld, Ansehen, Umwelt, Ziele), kulturelle Ausprägung des Arguments (IND, KOL) und kultureller Hintergrund der Nutzer (IND, KOL). Zusätzlich handelte es ich bei den ersten drei Variablen um Within-Subjects-Variablen, da alle Teilnehmer der Studie alle Kombinationen aus Empfehlungen und Argumenten bewerteten. Der kulturelle Hintergrund der Nutzer war dagegen ein Between-Subjects-Faktor, da alle Teilnehmer nur Mitglied einer kulturellen Gruppe sein konnten. Der Mixed-ANOVA-Test untersuchte sowohl den Effekt einzelner Variablen auf die Überzeugungskraft als auch Interaktionseffekte zwischen den Variablen.

Effekte einzelner Variablen Bei der Analyse der Effekte der einzelnen Variablen konnten für das Empfehlungsthema ($F(2, 182) = 3{,}99$; $p < 0{,}05$), das Argumentthema ($F(2{,}71, 246{,}20) = 25{,}86$; $p < 0{,}001$) sowie für die kulturelle Ausprägung der Argumente ($F(1, 91) = 16{,}58$; $p < 0{,}01$) signifikante Einflüsse festgestellt werden. Der kulturelle Hintergrund der Teilnehmer hatte alleine keinen Einfluss auf die wahrgenommene Überzeugungskraft der Empfehlungstexte. Die Ergebnisse der paarweisen Vergleiche (Bonferroni-Post-Hoc-Test) innerhalb der Variablen sind in Tabelle 5.6 zu sehen. Bei den Empfehlungsthemen schnitten Stromspartipps signifikant besser ab als Empfehlungen zur Einsparung von Heizenergie. Argumente, die sich mit den Themen Geldsparen und Umweltschutz beschäftigten, wurden als signifikant überzeugender wahrgenommen als die anderen Argumente. Das Thema Geldsparen war allerdings auch signifikant überzeugender als das Thema Umweltschutz. Zu guter Letzt war die Überzeugungskraft individualistisch geprägter Argumente signifikant höher als die der kollektivistisch geprägten Argumente.

Tabelle 5.6: Paarweiser Vergleich der einzelnen Varianten der Variablen „Thema der Empfehlung“, „Thema des Arguments“ und „kulturelle Ausprägung der Argumente“

	Mittelwert	Standardabweichung	signifikant besser als
Thema der Empfehlung			
Strom	3,26	0,09	Heizung*
Heizung	3,10	0,10	
Benzin	3,16	0,10	
Thema des Arguments			
Geld	3,71	0,08	Ansehen***, Umwelt***, Ziele***
Ansehen	2,83	0,12	
Umwelt	3,21	0,12	Ansehen**, Ziele*
Ziele	2,95	0,12	
Kulturelle Ausprägung der Argumente			
IND	3,25	0,09	COL***
KOL	3,1	0,09	
*signifikant mit p <0.05; **signifikant mit p <0.01; ***signifikant mit p <0.001			

Interaktionseffekte zwischen Variablen Neben den signifikanten Haupteffekten ergab der Mixed ANOVA-Test auch signifikante Interaktionseffekte zwischen manchen der Variablen. Es gab einen Effekt zwischen den Themen der Empfehlungen und der Argumente ($F(4{,}038, 367{,}449) = 11{,}009$; $p < 0{,}001$), zwischen dem Thema und der kulturellen Ausprägung der Argumente ($F(1{,}705, 155{,}121) = 54{,}530$; $p < 0{,}001$) sowie zwischen dem Empfehlungsthema und dem Thema und der kulturellen Ausprägung der Argumente ($F(4{,}973, 452{,}569) = 3{,}988$; $p < 0{,}01$).

Betrachtet man Abbildung 5.6, so sieht man, dass die Überzeugungskraft eines Argumentthemas vom jeweiligen Thema der Empfehlung abhing.

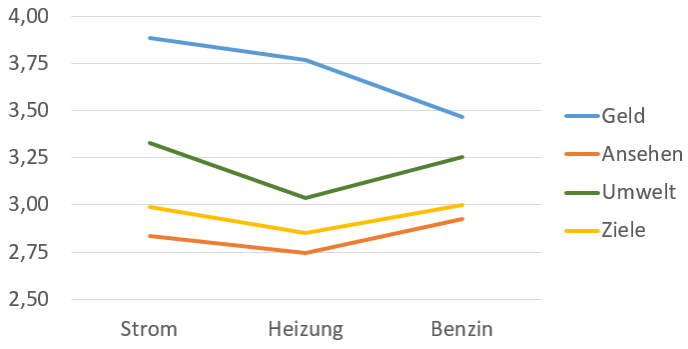

Abbildung 5.6: Geschätzte durchschnittliche Überzeugungskraft von Empfehlungen der Themen Strom, Heizung und Benzin mit Argumenten der Themen Geld, Ansehen, Umwelt und Ziele

Über alle Empfehlungsthemen hinweg waren finanzielle Ersparnisse die überzeugendsten Argumente. Bei Empfehlungen zum Benzinsparen war der Unterschied

allerdings weniger stark ausgeprägt, da die Überzeugungskraft finanzieller Argumente weniger stark ausfiel. Argumente zum Schutz und Erhalt der Umwelt wurden bei Empfehlungen zur Ersparnis von Heizenergie weniger überzeugend wahrgenommen als bei anderen Empfehlungen. Außerdem war ihre Überzeugungskraft bei Empfehlungen zum Benzinsparen nur etwas geringer als bei finanziellen Argumenten. Argumente bezüglich des eigenen Ansehens und bezüglich eigenen bzw. kollektiven Zielen schnitten bei allen Empfehlungsthemen in etwa gleich schlecht ab.

Der Interaktionseffekt zwischen dem Thema und der kulturellen Ausprägung der Argumente ist in Abbildung 5.7 dargestellt. Argument hinsichtlich finanzieller Vorteile waren viel überzeugender, wenn sie eine individualistische Ausprägung hatten. Argument, die auf Energiebilanzen oder Emissionsziele abzielen, hatten dagegen in der kollektivistischen Variante eine stärkere Überzeugungskraft. Für die beiden anderen Argumentthemen machte die kulturelle Ausprägung keinen Unterschied.

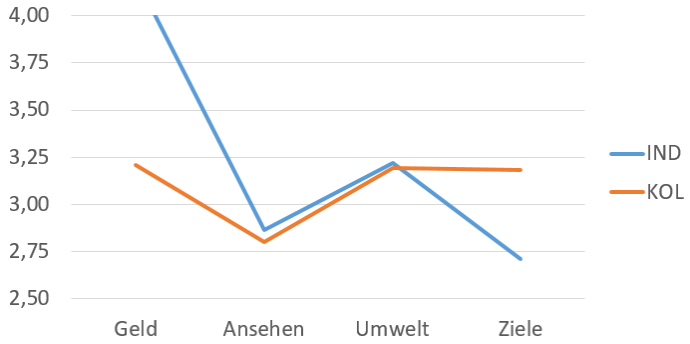

Abbildung 5.7: Geschätzte durchschnittliche Überzeugungskraft von Argumenten der Themen Geld, Ansehen, Umwelt und Ziele mit individualistischer und kollektivistischer Ausrichtung

Abbildung 5.8 zeigt den signifikanten Interaktionseffekt zwischen den Themen der Empfehlungen und der Argumente sowie der kulturellen Ausprägung der Argumente. Größtenteils bestätigen sich die bisherigen Erkenntnisse. Argumente für finanzielle Vorteile wiesen für alle Empfehlungsthemen die größte Überzeugungskraft auf. Diese Überzeugungskraft nahm allerdings von Stromspartipps über Tipps zur Einsparung von Heizenergie bis hin zu Empfehlungen für die Anschaffung oder Nutzung umweltfreundlicher Fortbewegungsmittel ab. Alle anderen Argumente wiesen unabhängig von der kulturellen Ausprägung und dem Empfehlungsthema weniger starke Schwankungen auf und hatten nur eine mittelmäßige Überzeugungskraft.

Interaktionseffekte mit kulturellem Hintergrund Mit Hilfe des Mixed ANOVA-Tests konnten auch signifikante Interaktionseffekte hinsichtlich des kulturellen Hintergrunds der Nutzer aufgedeckt werden. Die Effekte bestanden zum einen zwischen dem Argumentthema und der Kultur ($F(2{,}705,\ 246{,}196)=8{,}344$; $p<0{,}001$) und zum anderen zwischen den Themen der Empfehlungen und der Argumente sowie der Kultur ($F(4{,}038,\ 367{,}449)=3{,}807$; $p<0{,}01$).

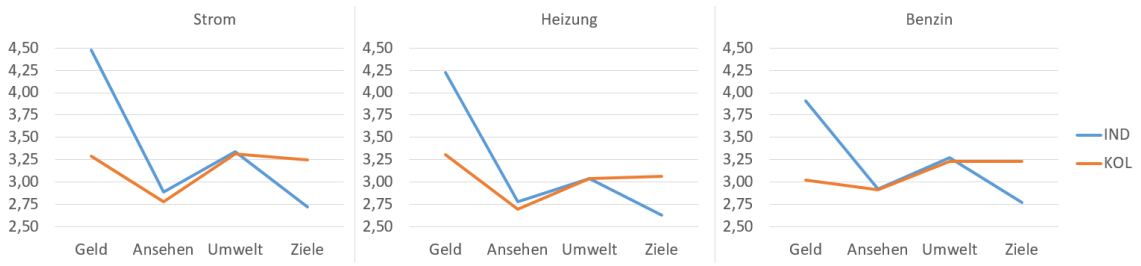

Abbildung 5.8: Geschätzte durchschnittliche Überzeugungskraft von Argumenten der Themen Geld, Ansehen, Umwelt und Ziele und einer individualistischen und kollektivistischen Ausrichtung für die Empfehlungsthemen Strom, Heizung und Benzin

In Abbildung 5.9 ist zu sehen, dass Mitglieder kollektivistischer Kulturen finanzielle Argumente mit Abstand am überzeugendsten fanden. Andere Argumente wurden nur als mittelmäßig überzeugend wahrgenommen. In individualistischen Kulturen wirkten finanzielle Argumente etwas weniger überzeugend. Dafür haben Argumente zum Schutz und Erhalt der Umwelt und Argumente, die ein gesteigertes Ansehen versprechen, in diesen Kulturen eine höhere Überzeugungskraft als in kollektivistischen Kulturen erreicht. Der Schutz der Umwelt stellte im Vergleich zu Geldersparnissen sogar ein ähnlich überzeugendes Argument dar.

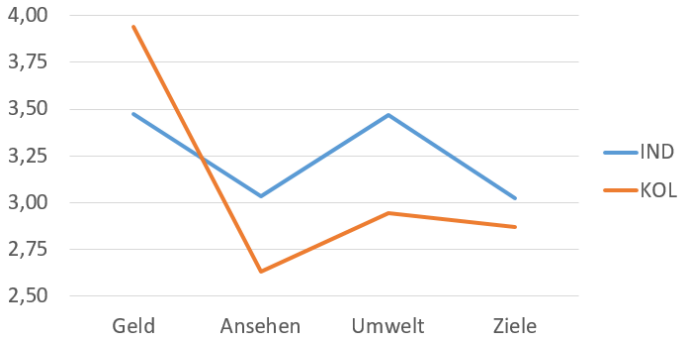

Abbildung 5.9: Geschätzte durchschnittliche Überzeugungskraft von Argumenten der Themen Geld, Ansehen, Umwelt und Ziele in individualistischen und kollektivistischen Kulturen

Die Analyse des Interaktionseffekts zwischen Empfehlungsthema, Argumentthema und kulturellem Hintergrund der Studienteilnehmer bestätigte, dass die Teilnehmer kollektivistischer Kulturen finanziellen Argumenten eine viel stärkere Überzeugungskraft zusprachen als anderen Argumenten, siehe Abbildung 5.10. Dieser Effekt war unabhängig vom Thema der Empfehlung. Bei individualistischen Kulturen war dies nicht der Fall. Das Thema Umwelt spielte generell eine ähnlich große Rolle wie Geldersparnisse. Bei Empfehlungen zum Benzinsparen stellte der Schutz und Erhalt der Umwelt sogar das überzeugendste Argument dar. Finanzielle Vorteile hatten hier ähnlich wenig Überzeugungskraft wie die restlichen Argumente.

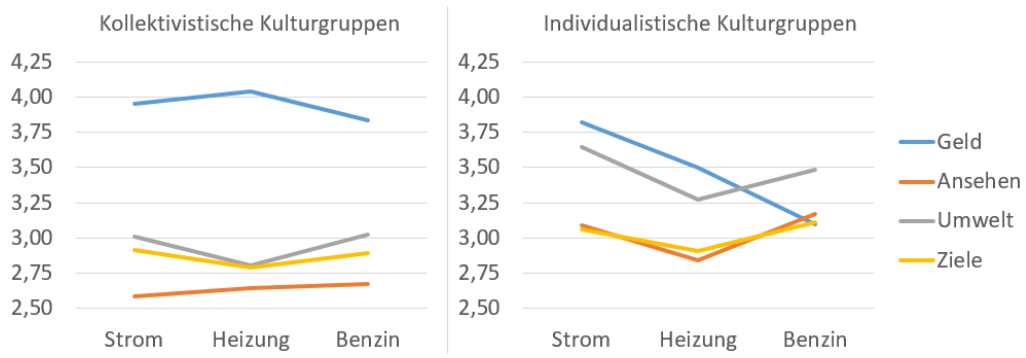

Abbildung 5.10: Geschätzte mittlere Überzeugungskraft von Empfehlungen der Themen Strom, Heizung und Benzin mit Argumenten der Themen Geld, Ansehen, Umwelt auf individualistische und kollektivistische Kulturen

Diskussion Der zweite Teil der Onlinestudie zeigte, dass eine kulturbasierte Auswahl von Argumenten für Energiesparmaßnahmen durchaus die Überzeugungskraft einer Empfehlung steigern kann. Die Begründungen vieler Studienteilnehmer für ihre abgegebenen Bewertungen unterstrichen die Erkenntnisse der statistischen Analyse und halfen beim Verständnis der kulturellen Unterschiede.

Sehr viele Aussagen der Teilnehmer zeigten die Wichtigkeit finanzieller Vorteile auf. Unabhängig von der jeweiligen Kultur wiederholten sich Aussagen wie „Saving money is always good.“ oder „Money rules the world.“ Zwei Teilnehmer aus Neuseeland schrieben „It is easier to convince people when reasoning it with economy and health advantages.„ bzw. „Saving costs is a larger motivation. Combining saving money with reducing emissions is best (in my opinion).“. Eine Italienerin war der Meinung, dass durchschnittliche Menschen relativ egoistisch sind und hauptsächlich durch finanzielle Gründe motiviert werden. Passend dazu schrieb eine US-Amerikanerin: „Saving money allows people to live within their means.„ Ein Spanier erklärt seine unterschiedliche Meinung zu den Themen Geld und Ansehen: „People in my culture are very concerned about spending money unnecessarily, and do not care what others think of how environmentally conscious they are.“

Dass speziell die Teilnehmer kollektivistischer Kulturen finanzielle Argumente überzeugender wahrnahmen als andere Argumente, unterstrichen die folgenden Aussagen griechischer (GRC) und ägyptischer (EGY) Teilnehmer: „The financial aspect is a very important factor in my country. (GRC)“, „In my culture money and looks talk. So if the reason involves how people perceive you or money its highly persuasive. (EGY)“, „Egyptians are also mostly above poverty threshold. So money saving arguments are very good. (EGY)“, „Money is the most important motive in a society with a majority of less educated. Some people of high standard levels have environmental awareness. (EGY)“

Die letzte Aussage ist bereits ein gutes Beispiel für die geringe Überzeugungskraft von Argumenten hinsichtlich des Schutzes und Erhalts der Umwelt in kollektivistisch geprägten Kulturen. Sie lässt Umweltbewusstsein wie eine Art Luxus für höher gestellte Menschen erscheinen. Auch Aussagen russischer Teilnehmer bestätigten, dass

der Schutz und Erhalt der Umwelt im eher kollektivistisch geprägten Russland kein motivierendes Argument darstellt: „In Russia people do not think about the environment.“. Hier scheint die Wahrnehmung umweltbewussten Verhaltens zumindest teilweise sogar eher als Schwäche ausgelegt zu werden: „The use of public transport means you do not have your own car - it's bad for your image. Most people buy a car (or e.g. expensive phones, clothes) which are too expensive and not necessary for these people, just because it is important for the image.“

Gegenteilige Aussagen zur Motivation durch finanzielle Argumente und Argumente bezüglich des Umweltschutzes konnten bei Teilnehmern aus individualistischen Kulturen gefunden werden: „Every little step is needed to keep our planet that beautiful for the next generations (DEU)“, „Using public transport to save money suggests you can't provide for yourself or your family (GBR)“, „ Many younger people do not anticipate having children (or grandchildren), but are concerned about the future of the world for people in general. (NZL)“, „Most people don't care about either their personal energy balance or the national emissions targets. Most do care about their impacts on costs, both financial and environmental. (NZL)“

Auch die Präferenz für individualistisch geprägte Argumente wurde durch Aussagen der Studienteilnehmer hervorgehoben: „Personal savings are the most motivating arguments. (DEU)“, „Average people definitely care about their own children, but might not be able to think the same for children of others (future generations) (ITA)“, „I think people are persuaded by things that directly impact them and have clear tangible effects in the near future. (NZL)“

Dass nationale Emissionsziele wenig überzeugend waren, lag den Aussagen einiger Teilnehmer zufolge daran, dass sie im präsentierten Argument „zu abstrakt (DEU)“ erwähnt wurden (z.B. „Purely mentioning emissions makes people think that it wouldn't really make a difference and does not motivate people. (NZL)“). Außerdem waren einige Personen der Meinung, dass nationale Ziele für normale Bürger weniger interessant sind: „People do not know enough about the national emmission targets. (ITA)„, „National emission targets aren't a big deal here. (NZL)“

Weitere Kommentare bestätigten allerdings auch, dass die Argumentauswahl nicht allein basierend auf einer Einordnung der Nutzer hinsichtlich Hofstedes Kulturdimensionen durchgeführt werden kann. Gerade im Rahmen der Empfehlung, häufiger zu Fuß zu gehen oder mit dem Rad zu fahren, forderten viele Teilnehmer aus individualistischen Kulturen themenbezogene Argumente: „For health and exercise related energy saving tips using the personal benefits of the exercise is probably a better way to promote the behaviour. (GBR)“, „Personal fitness would be a bigger incentive in New Zealand (NZL)“, „What about including being fit and more sexually attractive or looking very healthy and active without having to go to the gym that is too difficult and costly? (ITA)“, „Health and sports could be a good reason. (ESP)“, „I would add reasons about calorie and losing weight, or even building connections to the local community, or having time together with a loved one. (USA)“

Auch individuelle Ziele sollten laut der Aussage eines Deutschen nicht vernachlässigt werden. Um die Wirkung überzeugender Argumente auf Dauer hoch zu halten,

müssten diese Ziele allerdings detaillierter beschrieben werden und Ziele adressieren, die die Nutzer sich im System gesetzt haben.

5.1.4 Zusammenfassung

Dass ein kultureller Einfluss auf die Überzeugungskraft von Argumenten für Energiesparempfehlungen besteht, konnte vor allem mit Hilfe des zweiten Teils der Onlinestudie gezeigt werden (Hypothese 1). Allerdings konnten die Unterschiede nicht ohne Weiteres auf die Beschreibungen der Kulturen in Hofstedes Kulturmodell übertragen werden (Hypothese 2).

Es deutet sich an, dass so manche Kritik an Hofstedes Modell recht behalten sollte. Ein Kritikpunkt ist, dass die zur Ermittlung der kulturellen Unterschiede befragten Personen hauptsächlich IBM-Mitarbeiter waren und somit nicht repräsentativ genug waren, um Aussagen über ganze Kulturen zu treffen [McSweeney, 2002]. Außerdem ist einigen Wissenschaftlern das Konzept einer nationalen Kultur mit homogenen Menschen zu starr [McSweeney, 2002, Steinmetz, 1999]. Khaled [Khaled et al., 2006] beschrieb die neuseeländische Kultur zum Beispiel als stark gemischt. Ein Großteil der Population sei zwar europäischen Ursprungs. Allerdings bilden auch Maori und Menschen aus dem pazifischen Raum und Asien einen wichtigen Teil der Bevölkerung. Es leben also sowohl individualistisch als auch kollektivistisch geprägte Menschen in Neuseeland. In Hofstedes Kulturmodell wird dagegen allen Neuseeländern ein relativ stark ausgeprägter Individualismus (Wert: 79) zugesprochen. Eine Argumentauswahl, die sich nur auf Hofstedes kultureller Einschätzung des Herkunftsland der Nutzer bezieht, würde demzufolge seine Auswahl häufig basierend auf falschen Annahmen über die Werte und Bedürfnisse der individuellen Nutzer treffen.

Dass individuelle Präferenzen und Ziele bei der Argumentauswahl von Bedeutung sind, wurde auch dadurch deutlich, dass die Überzeugungskraft beliebter Argumentthemen wie *Geldsparen*, *Energiesparen* und *Umweltschutz* laut Aussagen der Studienteilnehmer mehr vom persönlichen Hintergrund der Nutzer, als von ihrer kulturellen Herkunft abhing. Eventuell könnte die Berücksichtigung der Energiekultur auch bei der Auswahl von Argumenten einen guten Beitrag leisten. Während Geld ein Motivationsfaktor für weniger umweltbewusste Menschen sein könnte, könnte der Schutz der Umwelt ein wichtiger Faktor für bereits umweltbewusst handelnde Personen sein. Allerdings sollten in beiden Fällen feinere Details und personalisierte Informationen wie zum Beispiel genaue Ersparnis erwähnt werden, um die Erklärungen nicht monoton und damit langfristig unwichtig werden zu lassen.

Wichtig anzumerken ist, dass in dieser Arbeit nur der Effekt einzelner Argumente verglichen wurde. Wie zu Beginn des Kapitels beschrieben wurde, sollten Empfehlungstexte allerdings eine Kombination zweier Argumente enthalten und weitere optionale Argumente und Informationen bereithalten. Kann man allerdings die Überzeugungskraft einzelner Argumente für individuelle Nutzer gut einschätzen, können wiederum bekannte Strategien zur Kombination von Argumenten eingesetzt werden.

5.2 Höflichkeitsstrategien in Empfehlungstexten

Höflichkeit wird als aufmerksames und rücksichtsvolles Verhalten gegenüber Anderen verstanden. Die individuellen Erwartungen an höfliches Verhalten hängen stark von den Werten und gewohnten Umgangsformen der Konversationspartner sowie der aktuellen Situation ab. Eigentlich höflich gemeintes Verhalten kann also je nach Konversationspartner und Situation auch falsch aufgefasst werden. Im Idealfall kann Höflichkeit allerdings zur Entwicklung von Solidarität, Vertrautheit und einer emotionalen Bindungen zwischen Interaktionspartnern beitragen [Cassell und Bickmore, 2003, Svennevig, 1999].

In der HCI wurde Höflichkeit bisher meist im Zusammenhang mit dem Verhalten virtueller und robotischer Agenten untersucht, siehe z.B. [Johnson et al., 2005, Srinivasan und Takayama, 2016]. Heerink [Heerink, 2010] zeigte allgemein, dass Roboter durch ein sozial intelligentes Verhalten wie den Ausdruck von Empathie oder das Führen einer angenehmen Konversation ihre soziale Präsenz und die Akzeptanz durch die Nutzer steigern können. Torrey und Kollegen [Torrey et al., 2013] fanden heraus, dass weniger direkt formulierte Anweisungen und Ratschläge Roboter weniger kontrollierend und dafür rücksichtsvoller und sympathischer wirken lassen. Allerdings konnten die Studienteilnehmer bei Torrey keine eigene Interaktion mit dem Roboter erleben, sondern bewerteten lediglich aufgezeichnete Szenen, in denen eine Person auf unterschiedliche Arten Anweisungen von einem Roboter erhielt. Durch solche Studien gewonnene Erkenntnisse sind jedoch nicht uneingeschränkt auf tatsächliche Interaktionen übertragbar [Strait et al., 2014]. Ein weiterer Nachteil bisheriger Studien zur Höflichkeit sozialer Roboter ist, dass meist nur direkte und indirekte Aussagen verglichen werden [Briggs und Scheutz, 2016]. Dabei gibt es, wie in diesem Kapitel beschrieben wird, viele weitere sprachliche Möglichkeiten die Höflichkeit einer Äußerung zu regulieren.

Forschungsfrage In dieser Dissertation wird die Frage beantwortet, ob unterschiedliche Höflichkeitsstrategien für die Formulierung von Empfehlungstexten genutzt werden können, um situativ spezifische Konversationsziele zu erreichen. Zu diesen Zielen zählen das Vermeiden von Gefühlen wie Bevormundung und Scham und der damit verbundene Aufbau einer guten Beziehung zu den Nutzern sowie eine gesteigerte Überzeugungskraft.

In Kapitel 5.2.1 werden zunächst bekannte Theorien und Strategien für Höflichkeit beschrieben. Anschließend wird eine Evaluation vorgestellt, mit der die Wahrnehmung von Höflichkeitsstrategien in Empfehlungstexten im Allgemeinen untersucht wurde, siehe Kapitel 5.2.2. Die Entwicklung eines Prototypen im CARE-Szenario sowie die Evaluation, die mit diesem Prototypen in einem lokalen Altersstift durchgeführt wurde, sind Bestandteil des Kapitels 5.2.3 sowie einer Veröffentlichung im Rahmen der PERSUASIVE-Konferenz 2016 [Hammer et al., 2016a]. Eine Diskussion der Ergebnisse in Kapitel 5.2.4 schließt dieses Kapitel ab.

5.2.1 Theorien und Strategien

Es gibt vier verbreitete Ansätze zur Charakterisierung von *Höflichkeit* [Fraser, 2001].

Social-Norm View Diese Theorie bezieht sich auf die Wahrung gesellschaftlicher Normen und der Erwartungen, die durch diese Normen zwischen Konversationspartnern entstehen. Durch förmliches und den Normen entsprechendes Verhalten, lässt man dieser Theorie zufolge dem Gegenüber die Aufmerksamkeit zu teil werden, die ihm oder ihr gebührt [Labov, 1984]. Ein Beispiel sind Grußformeln. Fremde Erwachsene und speziell Respektspersonen würde man beispielsweise nie mit einem „Hi“ oder „Servus“ begrüßen. Zwischen Bekannten und Freunden sind solche Grußformeln stattdessen weit verbreitet. Ein ähnliches Beispiel ist die Anrede mit „Du“ oder „Sie“. Im Hinblick auf den Einsatz in einem Empfehlungssystem bietet die Social-Norm View Theorie eine gute Grundlage für Normen, die generell befolgt werden sollten. Allerdings ist es bei dieser Form von Höflichkeit nicht vorgesehen situativ auch einmal absichtlich auf Höflichkeitsformen zu verzichten, um einen gewünschten Effekt (z.B. Pflegen der Beziehung oder Steigerung der Überzeugungskraft) zu erzielen [Erndl, 1998, Fraser, 2001]. Das oberste Ziel ist eine möglichst reibungslose Interaktion zwischen den Mitgliedern einer sozialen Gruppierung.

Conversational-Contract View Die Theorie der *Conversational Contracts* beschreibt die Konfliktvermeidung durch Regeln, auf die sich die Gesprächspartner zu Beginn eines Gesprächs abhängig von der Situation einigen [Fraser und Nolen, 1981]. Der jeweilige Sprecher wählt seine Verhaltensweise also nicht danach, welchen Zweck er für sich in der aktuellen Situation erfüllen möchte, sondern weil diese Verhaltensweisen in beidseitigem Einverständnis vorgegeben wurden. Dadurch ist die Conversational-Contract View Theorie ähnlich wie die Social-Norm View Theorie nur bedingt für den Einsatz in Systemen wie CARE oder SavER geeignet.

Conversational-Maxim View Anders als bei den ersten beiden Ansätzen gehen Sprecher dieser Höflichkeitstheorie folgend so vor, dass sie ihr Ziel - eine Kooperation mit dem jeweiligen Kooperationspartner - erreichen. Nach Grice [Grice, 1975] gibt es hierfür vier Maximen, an die man sich als Sprecher halten sollte. Sie gelten als Grundvoraussetzung für eine Kommunikation.

- Maxime der Quantität: Teile nicht mehr und nicht weniger Information als nötig mit.
- Maxime der Relation: Teile nur für die Situation relevante Informationen mit.
- Maxime der Qualität: Teile keine Informationen mit, über deren Richtigkeit du Zweifel hast. Informiere auch über die Wahrscheinlichkeit der Richtigkeit.
- Maxime der Modalität: Teile die Informationen so klar wie möglich mit und passe die Art und Weise an die jeweilige Situation an.

Das Maximen-Modell von Lakoff fasst die Maximen von Grice zu *Rules of Conversation* zusammen, die besagen, dass man sich klar und deutlich ausdrücken soll und ergänzt die folgenden *Rules of Politeness* [Lakoff, 1973]:

- Wahre Distanz zum Konversationspartner (z.B. Verwende Passivsätze, Nachnamen und Anrede. Frage bei Störungen und Einmischungen um Erlaubnis.)
- Gebe Freiraum für Entscheidungen (z.B. „Ich würde sagen, es ist langsam Zeit das Fenster zu schließen."; häufig in Verbindung mit Wahrung der Distanz)
- Sei freundlich und sorge dafür, dass sich der Konversationspartner wohl fühlt. (z.B. Bestätige durch Komplimente, Verwende Vor- oder Spitznamen)

Ein weiteres Maximen-basiertes Modell ist das Höflichkeitsprinzip-Modell [Leech, 1983]. Laut ihm steht die Höflichkeit als regulierendes Sprachmittel über den Maximen von Grice. Sie erfüllt den Zweck, ein soziales Gleichgewicht und eine freundliche Beziehung zwischen Konversationspartnern herzustellen. Das Modell ergänzt ebenfalls die Maximen von Grice durch weitere Maximen:

- Maxime des Taktes: Minimiere die Kosten und maximiere den Nutzen des Anderen
- Maxime der Großzügigkeit: Minimiere den eigenen Nutzen und maximiere den Nutzen des Anderen
- Maxime der Bestätigung: Minimiere Kritik und maximiere Lob am Anderen
- Maxime der Bescheidenheit: Minimiere eigenes Lob und maximiere Selbstkritik
- Maxime der Einigkeit: Minimiere Uneinigkeit und maximiere Einigkeit mit dem Anderen
- Maxime der Sympathie: Minimiere Antipathie und maximiere Sympathie zwischen dir und dem Anderen

Die Maximen-Modelle liefern eine gute Basis für höfliches Verhalten. Allerdings fehlen konkrete Strategien, die in Empfehlungstexten verwendet werden könnten.

Face-Saving View Der Face-Saving View und vor allem die Höflichkeitsstrategien von Brown und Levinson [Brown und Levinson, 1987] beschreiben ein zweckorientiertes Vorgehen. Sie bieten verschiedene Strategien, mit denen in unterschiedlichen Situationen eine Balance zwischen dem Ziel des Sprechers und der Gesichtswahrung des Konversationspartners erreicht werden kann. Fraser [Fraser, 2001] bezeichnete das Modell von Brown und Levinson als das beste und kompletteste Modell im Bezug auf linguistische Höflichkeit. Auch in der HCI fand es schon Beachtung. Johnson und Kollegen [Johnson et al., 2005] nutzten die Höflichkeitsstrategien zum Beispiel, um die Wahrnehmung des Verhaltens virtueller Agenten in unterschiedlichen Kulturen zu untersuchen.

Brown und Levinson [Brown und Levinson, 1987] definieren Höflichkeit als eine grundlegende Voraussetzung für die Kooperation zwischen Menschen. Sie bieten verschiedene sprachliche Strategien an, die Sprecher im Hinblick auf ein persönliches Ziel anwenden können, ohne unnötig grundlegende Bedürfnisse ihres Gegenüber zu verletzen. Diese Bedürfnisse sind die Gewährleistung der persönlichen Privatsphäre, Autonomie und Selbstverantwortung, das sog. *Negative Face*, und das Recht auf Anerkennung und der Wunsch, dass eigene Bedürfnisse von anderen Menschen ebenfalls als wünschenswert betrachtet werden, das sog. *Positive Face*. Handlungen, die diese Bedürfnisse bedrohen, werden als *Face Threatening Acts (FTAs)* bezeichnet. Sie können den Sprecher selbst aber auch den Zuhörer oder Adressaten betreffen.

Aus Sicht dieser Dissertation sind allerdings vor allem FTAs zu berücksichtigen, die Nutzer, also die Adressaten, betreffen. Proaktive Erinnerungen und Empfehlungen bedrohen die Autonomie und Selbstbestimmung der Nutzer. Außerdem werden in ihnen möglicherweise kontroverse Themen angesprochen und auf Missstände hingewiesen. Dies kann von den Nutzern als Kritik am eigenen Lebensstil und Verhalten wahrgenommen werden. Formuliert man die Empfehlungen höflicher, können diese Gefahren eventuell verringert werden. Allerdings kann eine erhöhte Höflichkeit auch die Überzeugungskraft einer Empfehlung abmildern und damit das Ziel, durch die empfohlene Maßnahme das Wohlbefinden zu fördern, gefährden.

Die Balance zwischen der Verfolgung des eigenen Ziels und der Wahrung des Gesichts des Gegenübers ist ein zentraler Punkt der Höflichkeitsstrategien von Brown und Levinson, die im nächsten Abschnitt beschrieben werden. Sie berücksichtigen implizit eine bewusstes und zweckgebundenes in Kauf nehmen einer FTA, wenn sie für das Erreichen des übergeordneten Ziels unausweichlich ist [Brown und Levinson, 1987]. Für CARE und SavER bedeutet das, dass situativ durchaus der Grad der Höflichkeit einer Formulierung reduziert werden könnte, um einer Zielperson indirekt klar zu machen, dass es notwendig ist, aktiv zu werden oder der aktuellen Empfehlung zu folgen.

Höflichkeitsstrategien nach Brown und Levinson [Brown und Levinson, 1987] Um abschätzen zu können, wie höflich eine Aussage formuliert sein sollte, muss ein Sprecher abhängig vom Konversationspartner und der Situation verschiedene Kriterien analysieren.

Ein ausschlaggebendes Kriterium ist die soziale Distanz zwischen den Konversationspartnern. Sie beschreibt, ob sich Sprecher (S) und Hörer (H) als gleichwertig ansehen und gegebenenfalls näher stehen, ob eine gewisse Distanz zwischen den beiden besteht oder ob gar einer der beiden die andere Person zum Beispiel aufgrund einer höheren gesellschaftlichen Position dominiert. Eine typische Auswirkung sozialer Distanz im Deutschen ist die Ansprache mit „Sie“ oder „Du“.

Das zweite wichtige Kriterium ist die relative Macht, die H auf S ausübt. Diese Macht kann materiellen, körperlichen sowie metaphysischen Ursprungs sein. Sie beschreibt, wie sehr H seine eigenen Pläne und seine Selbsteinschätzung S auferlegen kann und zwar auf Kosten der Pläne und Selbsteinschätzung von S.

Zu guter Letzt spielt auch die gesellschaftliche Einschätzung verschiedener FTAs eine wichtige Rolle. Abhängig von der jeweiligen Kultur und Situation besteht eine Art Ranking hinsichtlich der Zumutbarkeit von FTAs.

Um die Gewichtigkeit Wx einer FTA anhand der genannten Kriterien berechnen zu können, entwickelten Brown und Levinson die folgende Gleichung:

$$Wx = Distanz(S, H) + Macht(H, S) + Zumutbarkeit(FTA) \tag{5.1}$$

Abhängig von Wx kann eine geeignete Höflichkeitsstrategie ausgewählt werden. In Abbildung 5.11 sind vier Strategien, die Brown und Levinson in ihrer Arbeit mit vielfältigen Formulierungsmöglichkeiten präsentierten, sowie die Umstände in denen sie verfolgt werden sollten.

Besteht nur eine sehr geringe Gefahr eines Gesichtsverlusts des Anderen, so kann eine *direkte, unverblümte Äußerung* gewählt werden. In diesem Fall kann der FTA-Aspekt einer Aussage vernachlässigt werden. Als Sprecher übernimmt man die Verantwortung für den Kern der Aussage und versucht, möglichst eindeutig seine Information zu transportieren. Ein Beispiel wäre „Das Licht im Schlafzimmer ist immer noch an. Schalte es aus.“ oder „In Zukunft solltest du daran denken, das Licht auszuschalten, wenn du es nicht mehr brauchst.“

Ist man sich des FTA-Aspekts bewusst und möchte sozusagen Schadensbegrenzung betreiben, so bieten sich zwei Strategien an, die versuchen, gezielt einen Teil der Bedürfnisse des Anderen zu befriedigen: *positive* und *negative Höflichkeit*

Soll das Bedürfnis nach Anerkennung und Berücksichtigung der eigenen Wünsche und Werte befriedigt werden, bedient man sich der positiven Höflichkeit. Man versucht Gemeinsamkeiten hervorzuheben und sich dem Gegenüber anzunähern (vgl. soziale Distanz). Häufig wird zum Beispiel das Wort „wir“ benutzt. Beispielhafte Formulierungen sind „Wir brauchen das Licht im Schlafzimmer nicht mehr. Lass es uns ausschalten.“, „Ich weiß, dass es schwer ist, seinen inneren Schweinehund zu überwinden. Es würde sich aber bestimmt gut anfühlen, heute mit dem Fahrrad statt mit dem Auto zur Arbeit zu fahren.“ oder „Unsere Nachbarn fahren regelmäßig mit dem Fahrrad zur Arbeit. Wir sollten es auch einmal ausprobieren.“

Die negative Höflichkeit versucht eine Einschränkung der Autonomie des Gegenübers zu vermeiden. Dies wird erreicht, indem man Formulierungen verwendet, die einen gewissen Handlungs- oder Entscheidungsspielraum gewähren. Beispiele sind „Würde es dir etwas ausmachen das Licht im Schlafzimmer auszuschalten?“ oder „Du könntest bestimmt auch mit dem Fahrrad fahren, oder nicht?“, aber auch „Ich würde gerne das Licht im Schlafzimmer ausschalten, wenn es dir nichts ausmacht.“

Muss eine FTA ausgeführt werden und besteht ein hohes Risiko, dass der Gegenüber sein Gesicht verlieren könnte, so bleibt noch eine letzte Strategie. Durch eine *Äußerung mit einem hohen Maß an Mehrdeutigkeit* kann versucht werden, *indirekt* auf das tatsächliche Ziel der Aussage hinzuweisen. Sollte die Andeutung dennoch verstanden werden, schiebt man in gewisser Weise die Verantwortung für eine FTA von sich. Es wird dann zwar das Bedürfnis nach Bestätigung und Anerkennung verletzt. Die Selbstbestimmung bleibt allerdings weitestgehend erhalten, da die Möglichkeit

besteht, die Andeutung zu ignorieren. Beispiele sind „Das Licht im Schlafzimmer ist immer noch an.“ oder „Es ist gut, wenn man statt mit dem Auto zu fahren das Fahrrad verwendet.“

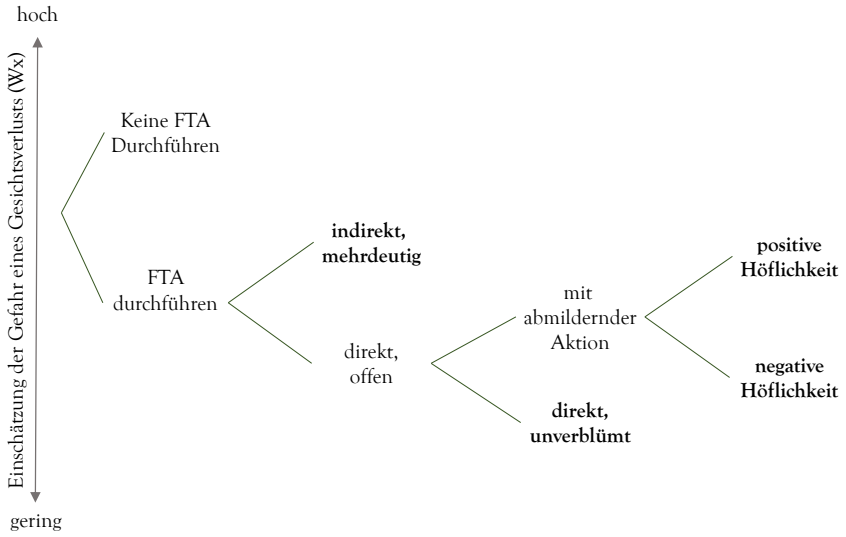

Abbildung 5.11: Höflichkeitsstrategien und die Umstände ihrer Auswahl nach [Brown und Levinson, 1987]

5.2.2 Evaluation der Wahrnehmung von Höflichkeitsstrategien

Um erste Erkenntnisse über die Wahrnehmung verschiedener Höflichkeitsstrategien im Kontext von Empfehlungssystemen zu erhalten, wurde eine Befragung mit jüngeren Erwachsenen durchgeführt. Hierfür wurden acht Arten von Formulierungen betrachtet, die auf den Höflichkeitsstrategien von Brown und Levinson [Brown und Levinson, 1987] beruhten und bereits von Johnson und Kollegen [Johnson et al., 2005] im Bezug auf Empfehlungen durch pädagogische Agenten untersucht wurden. Eine Übersicht der Strategien und beispielhafte Formulierungen der Empfehlung „Machen Sie die Fenster auf.“ sind in Tabelle 5.7 zusammengefasst. Das Set enthielt *direkte Kommandos* und *indirekte und mehrdeutige Äußerungen*, die als Ziel des Systems formuliert wurden. Strategien der positiven Höflichkeit waren durch *Anfragen und Bitten* und *Formulierungen als gemeinsames Ziel* vertreten. Zu den negativ höflichen Formulierungen gehörten *indirekte Andeutungen*, *Fragen*, *Andeutungen bzgl. der Absichten des Nutzers* und *Sokratische Hinweise*.

Hypothesen Das Ziel der Evaluation war die Untersuchung zweier Hypothesen für den Einsatz von Höflichkeitsstrategien in Empfehlungssystemen:

1. Die Art der verwendeten Höflichkeitsstrategien beeinflusst die wahrgenommene Höflichkeit einer Empfehlung.

2. Die Art der verwendeten Höflichkeitsstrategien beeinflusst die wahrgenommene Überzeugungskraft einer Empfehlung.

Tabelle 5.7: Höflichkeitsstrategien nach [Johnson et al., 2005] und beispielhafte Formulierungen für die Empfehlung „Machen Sie die Fenster auf“

Höflichkeitsstrategie	**Formulierung**	**Einschätzung**
Direktes Kommando	Machen Sie kurz die Fenster auf.	Direkt, unverblümt
Indirekte Andeutung	Die Sensoren im Raum sehen vor, dass Sie kurz die Fenster aufmachen.	Negative Höflichkeit
Fragen	Wie wäre es, wenn Sie kurz die Fenster aufmachen würden?	Negative Höflichkeit
Andeutungen bzgl. der Absichten des Nutzers	Sie möchten bestimmt kurz die Fenster aufmachen.	Negative Höflichkeit
Sokratischer Hinweis	Haben Sie daran gedacht kurz die Fenster aufzumachen?	Negative Höflichkeit
Anfrage/Bitte	Ich hätte gern, dass Sie kurz die Fenster aufmachen.	Positive Höflichkeit
Formulierungen als gemeinsames Ziel	Wir sollten kurz die Fenster aufmachen.	Positive Höflichkeit
Formulierungen als Ziel des Systems	Ich würde kurz die Fenster aufmachen.	Indirekt, mehrdeutig

Im Falle der Gültigkeit der Hypothesen wäre es möglich, situativ Höflichkeitsstrategien für Empfehlungstexte in assistierenden Empfehlungssystemen auszuwählen, um einen erwünschten Grad der Überzeugungskraft und Höflichkeit zu erreichen.

Durchführung Nach einer kurze Einführung erhielten die Studienteilnehmer einen Fragebogen mit demographischen Fragen. Anschließend wurden ihnen die Empfehlungen „Trinken Sie etwas Wasser.“, „Machen Sie kurz die Fenster auf.“ und „Gehen Sie etwas spazieren.“ jeweils in acht auf den Höflichkeitsstrategien basierenden Formulierungen (siehe Beispiel in Tabelle 5.7) in Textform zur Bewertung vorgelegt. Durch die verschiedenen Empfehlungen sollte die Gefahr verringert werden, dass die Bewertungen der Teilnehmer durch die Präferenzen für eine bestimmte Empfehlung beeinflusst werden. Um auch die Gefahr von Reihenfolgeeffekten zu mindern, wurden die Teilnehmer in unterschiedlichen Reihenfolgen mit den Empfehlungen und Formulierungen konfrontiert. Die Bewertungen der einzelnen Formulierungen waren im Hinblick auf die beiden zu überprüfenden Hypothesen zweigeteilt. Sie erfolgten mit Hilfe 7-stufiger Likert-Skalen von 1 = „überhaupt nicht höflich“ bzw. „überhaupt nicht überzeugend“ bis 7 = „sehr höflich“ bzw. „sehr überzeugend.“

Ergebnisse An der Befragung beteiligten sich insgesamt fünf weibliche und zehn männliche Teilnehmern im Alter von 21 bis 50. Die Altersverteilung ist in Abbildung 5.12 dargestellt. Alle gaben Deutsch als ihre Muttersprache an.

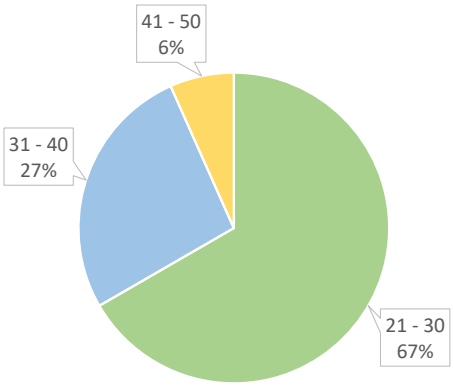

Abbildung 5.12: Altersverteilung der Teilnehmer der Befragung

Die durchschnittlichen Bewertungen, siehe Abbildung 5.13, zeigen, dass Fragen als am höflichsten wahrgenommen wurden. Formulierungen als gemeinsame Ziele oder als Ziele des Systems, aber auch Anfragen und Bitten wurden ebenfalls als höflich wahrgenommen. Im Gegensatz dazu wurden direkte Kommandos als am wenigsten höflich bewertet. Allerdings wurden direkte Kommandos dafür als ähnlich überzeugend wahrgenommen wie Fragen. Bezüglich dieses Kriteriums schlossen sokratische Hinweise und Andeutungen bzgl. der Absichten des Nutzers am schlechtesten ab. Strategien, die sowohl als höflich als auch als überzeugend bewertet wurden, waren Formulierungen als gemeinsames Ziel und Anfragen und Bitten.

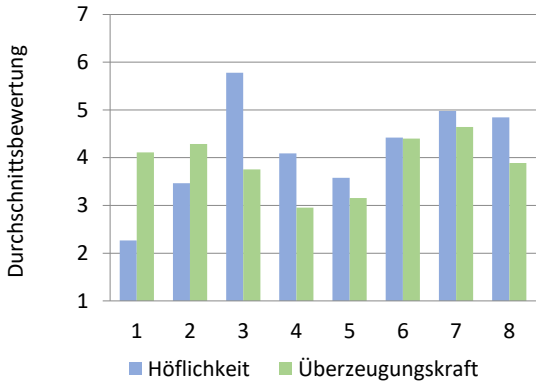

Abbildung 5.13: Durchschnittliche Bewertung der Höflichkeit und Überzeugungskraft der Formulierungen. 1: Direkte Kommandos, 2: Indirekte Andeutungen, 3: Fragen, 4: Andeutungen bzgl. der Absichten des Nutzers, 5: Sokratischer Hinweis, 6: Anfrage/Bitte, 7: Formulierungen als gemeinsames Ziel, 8: Formulierungen als Ziel des Systems

Ein ANOVA-Test für Messwiederholungen bestätigte die beiden aufgestellten Hypothesen. Sowohl für die wahrgenommene Höflichkeit ($F(7, 308) = 37.82, p < 0.001$) als auch für die wahrgenommene Überzeugungskraft ($F(5.25, 231.05) = 8.14, p < 0.001$) zeigten sich höchst signifikante Unterschiede. Der anschließende paarweise Vergleich mittels eines Bonferroni Post-Hoc Tests bestätigte nochmals die

Eindrücke der deskriptiven Analyse. Wie in Tabelle 5.8 zu sehen ist, wurden *Fragen* als signifikant höflicher alle anderen Formulierungen wahrgenommen. *Direkte Kommandos* dagegen wurden als signifikant unhöflicher als alle anderen Varianten bewertet. Des Weiteren zeigten die paarweisen Vergleiche zwar keine signifikanten Unterschiede zwischen den als überzeugend bewerteten Formulierungen. Sie zeigten allerdings, dass *Sokratische Hinweise* und insbesondere *Andeutungen bzgl. der Absichten des Nutzers* als signifikant weniger überzeugender wahrgenommen wurden als viele der anderen Höflichkeitsstrategien.

Tabelle 5.8: Wahrnehmung der Höflichkeitsstrategien (Abkürzungen: M = Mittelwert; SA = Standardabweichung; Sig = signifikant besser als)

	Höflichkeit			Überzeugungskraft		
	M	SA	Sig	M	SA	Sig
1. Direkte Kommandos	2.27	1.03		4.11	1.76	
2. Indirekte Andeutungen	3.47	1.41	1**	4.29	1.63	4**; 5*
3. Fragen	5.78	1.11	1,2,4,5,6,8*** 7*	3.76	1.32	4*
4. Andeutungen bzgl. der Absichten des Nutzers	4.09	1.41	1***	2.96	1.38	
5. Sokratische Hinweise	3.58	1.20	1***	3.16	1.51	
6. Anfragen/Bitten	4.42	1.60	1***	4.40	1.48	4*** 5**
7. Formulierungen als gemeinsames Ziel	4.98	1.22	1,2,5***	4.64	1.45	4*** 5**
8. Formulierungen als Ziel des Systems	4.84	1.11	1,2,5***	3.89	1.45	4*
*signifikant mit p <0.05; **signifikant mit p <0.01; ***signifikant mit p <0.001						

Diskussion Die Befragung zeigte, dass unterschiedliche Höflichkeitsstrategien nicht nur, wie bereits von Brown und Levinson [Brown und Levinson, 1987] und Johnson und Kollegen [Johnson et al., 2005] beschrieben, als unterschiedlich höflich, sondern auch als unterschiedlich überzeugend wahrgenommen werden können. Allerdings ist zu beachten, dass eine höfliche Formulierung nicht gleichzeitig auch als überzeugend wahrgenommen werden muss. Am deutlichsten zeigte dies der Vergleich von direkten Kommandos und Fragen. Während die Höflichkeit von Fragen signifikant besser bewertet wurde als die direkter Kommandos, waren direkte Kommandos sogar etwas überzeugender als Fragen, siehe Tabelle 5.8.

Basierend auf diesen Ergebnissen bietet sich für beratende Empfehlungssysteme die Möglichkeit an, situativ und abhängig von der Empfehlung die Balance zwischen Höflichkeit und Überzeugungskraft zu steuern. Grundsätzlich ist es das oberste Ziel, eine Formulierung zu wählen, die höflich und überzeugend ist. Laut der durchgeführten Befragungen eignen sich hierfür z.B. Formulierungen als gemeinsames Ziel oder

Anfragen und Bitten, die beide das Bedürfnis nach Anerkennung und Berücksichtigung der eigenen Wünsche und Werte (positive Höflichkeit) befriedigen. Besteht allerdings eine hohe Dringlichkeit, eine empfohlene Aktion durchzuführen - zum Beispiel, wenn durch ein sehr schlechtes Raumklima oder eine zu geringe Trinkmenge das Wohlbefinden der Nutzer gefährdet ist - könnte gerade der Verzicht auf ein gewisses Maß an Höflichkeit zu Gunsten einer stärken Überzeugungskraft hilfreich sein. So könnten sich die Nutzer auch durch die gewählte Formulierung der Dringlichkeit der Situation besser bewusst werden. Beispiele für geeignete Strategien sind direkte Kommandos oder indirekte Andeutungen, die den Nutzern keinen oder nur einen geringen Entscheidungsspielraum lassen. Allerdings kann dies auch auf Kosten des Verhältnisses zwischen Nutzer und System gehen. Deswegen wäre es eine weitere gute Strategie in Situationen, in denen es nur einen geringen Unterschied macht, ob eine Empfehlung angenommen wird oder nicht, eine höflichere Formulierung zu wählen, die allerdings nicht zwingend auch überzeugend sein muss. Durch Fragen oder Formulierungen als Ziel des Systems könnte das Verhältnis zwischen Nutzer und System und damit auch weitere zukünftige Interaktionen gefördert werden.

Da es sich bei dieser ersten Evaluation nur um eine textbasierte Befragung in einer Laborumgebung handelte, wurde zusätzlich eine realistischere Evaluation mit einem Prototypen für den Anwendungsfall CARE durchgeführt.

5.2.3 Evaluation im St.Jakobs Stift Augsburg

Das Problem, dass die Präsentation von Empfehlungen in beratenden Empfehlungssystemen dazuführen kann, dass sich Nutzer peinlich berührt und bevormundet fühlen, kann im CARE-Szenario verstärkt auftreten. Aufgrund körperlicher und mentaler Einschränkungen sind ältere Menschen häufig auf die Hilfe Anderer angewiesen und fühlen sich deswegen auch ohne ein Empfehlungssystem häufig minderwertig und bevormundet. Weist eine Empfehlung nun auch noch auf ihre Schwächen hin und versucht Vorschläge zu machen, wie ihre Situation verbessert werden könnte, so könnten sich diese Gefühle noch verstärken und letztendlich zu einer Ablehnung des Systems führen.

Aus diesem Grund war es nach den vielversprechenden Ergebnissen der ersten Evaluation von Höflichkeitsstrategien in Empfehlungstexten, siehe Kapitel 5.2.2, von besonderer Bedeutung, die Strategien auch auf ihre Wahrnehmung durch ältere Menschen hin zu untersuchen. Im Fokus standen dabei wiederum die Kriterien Höflichkeit und Überzeugungskraft und ihr Zusammenspiel.

Prototypische Umsetzung In dieser Evaluation sollte eine realistische, soziale Interaktion zwischen Nutzern und System simuliert werden. Hierfür wurde ein Prototyp mit einem sozialen, humanoiden Roboter entwickelt, der die Empfehlungen in den unterschiedlichen Varianten aussprechen sollte.

Empfehlungen Im Prototypen wurden dieselben Empfehlungen umgesetzt wie in der vorangegangenen Evaluation, siehe Kapitel 5.2.2: (1) „Trinken Sie etwas Wasser." (Formulierungen siehe Tabelle 5.9), (2) „Machen Sie kurz das Fenster auf" (Formulierungen siehe Tabelle 5.7) und (3) „Gehen Sie etwas spazieren." (Formulierungen analog zu (1) und (2)). Alle drei Aktionen sind insbesondere für das Wohlbefinden älterer Menschen von besonderer Wichtigkeit und werden leider allzu häufig vergessen oder vernachlässigt.

Tabelle 5.9: Höflichkeitsstrategien nach [Johnson et al., 2005] und beispielhafte Formulierungen für die Empfehlung „Trinken Sie etwas Wasser"

Höflichkeitsstrategie	**Formulierung**	**Einschätzung**
Direktes Kommando	Trinken Sie etwas Wasser.	Direkt, unverblümt
Indirekte Andeutung	Ihr Ernährungsplan sieht vor, dass Sie etwas Wasser trinken.	Negative Höflichkeit
Fragen	Wie wäre es, wenn Sie etwas Wasser trinken würden?	Negative Höflichkeit
Andeutungen bzgl. der Absichten des Nutzers	Sie möchten bestimmt etwas Wasser trinken.	Negative Höflichkeit
Sokratischer Hinweis	Haben Sie daran gedacht etwas Wasser zu trinken?	Negative Höflichkeit
Anfrage/Bitte	Ich hätte gern, dass Sie etwas Wasser trinken.	Positive Höflichkeit
Formulierungen als gemeinsames Ziel	Wir sollten kurz etwas Wasser trinken.	Positive Höflichkeit
Formulierungen als Ziel des Systems	Ich würde etwas Wasser trinken.	Indirekt, mehrdeutig

Sozialer Roboter Es wurde der Roboter Reeti der Firma Robopec[21] eingesetzt, siehe Abbildung 5.14. Er verfügt über bewegliche Augen, die vielfältiges Blickverhalten ermöglichen. Außerdem kann Reeti durch Animationen des Mundes, der Augenlider und der Ohren menschliche Mimik und Emotionen simulieren. Durch sein cartoonhaftes, außerirdisches bzw. fantastisches Aussehen wirkt Reeti allerdings nicht menschlich genug, um zu große Erwartungen an ein realistisches Verhalten zu wecken. Ein weiterer Vorteil für den Einsatz Reetis im Anwendungsfall CARE ist, dass er durch den unbeweglichen Körper, seine relativ geringe Größe ($44cm$) und seine einfarbig weiße Farbgebung in keinster Weise bedrohlich erscheint.

Zur Unterstützung seines non-verbalen Verhaltens kann Reeti die Backen in verschiedenen Farben aufleuchten lassen. Für die Sprachsynthese in Deutsch verwendet er die Text-to-Speech-Software Loquendo der Firma Nuance[22].

[21] http://www.reeti.fr

[22] http://www.nuance.com

Abbildung 5.14: Reeti - Ein sozialer, humanoider Roboter mit expressivem Kopf

Implementierung Die Studie wurde als Wizard-of-Oz-Studie durchgeführt. Über einfache Tastatureingaben konnte der Studienleiter zwischen den Formulierungen der Empfehlungen navigieren, sie wiederholen oder die Studie jederzeit mit einer abschließenden Aussage des Roboters beenden. Die Funktionalität zum Studienabbruch wurde integriert, da zu erwarten war, dass manche der Studienteilnehmer aufgrund ihres hohen Alters und damit verbundenen körperlichen oder geistigen Einschränkungen nicht die komplette Studie durchführen würden können. Es sollte allerdings sichergestellt werden, dass die Studie für jede Person mit positiven, abschließenden Worten und einem Dank durch Reeti beendet werden konnte.

Umgesetzt wurde die beschriebene Funktionalität mit einer halb-automatischen Dialogapplikation, die den von Mehlmann und Kollegen [Gebhard et al., 2012, Mehlmann et al., 2015] entwickelten *VisualSceneMaker* nutzte.

Um die Nutzer in ihrer Wahrnehmung der verbalen Aussagen des Roboters nicht zu beeinflussen, zeigte Reeti während des Sprechens immer eine neutrale Mimik und verzichtete abgesehen von den Lippenbewegungen auf jedwede weitere Animation. Um ihm aber dennoch etwas Leben einzuhauchen und ihn zugänglicher wirken zu lassen, führte Reeti zu Beginn der Studie und in den Pausen zwischen den Empfehlungen kleinere zufällige Bewegungen mit dem Kopf oder den Augenlidern aus.

Hypothesen Das Ziel der Studie war die Bestätigung der Ergebnisse aus der ersten Evaluation für Personen der CARE-Zielgruppe in einer zwar simulierten, aber dennoch realistischen Interaktion. Demzufolge lauteten die Hypothesen in dieser Studie ebenfalls:

1. Die Art der verwendeten Höflichkeitsstrategien beeinflusst die wahrgenommene Höflichkeit einer Empfehlung.

2. Die Art der verwendeten Höflichkeitsstrategien beeinflusst die wahrgenommene Überzeugungskraft einer Empfehlung.

Außerdem wurde die Akzeptanz des präsentierten Systems untersucht.

Durchführung Die Studie wurde in Kooperation mit dem St.Jakobs Stift in Augsburg durchgeführt. Mit Hilfe der Leitung des Altenheims wurden Senioren für die Studien angeworben, die kleinere körperliche und/oder geistige Einschränkungen hatten, aber immer noch dazu fähig waren, selbstständig große Teile ihres Alltags zu bewältigen. Zu ihnen zählten Bewohner des Altenheims und des betreuten Wohnens, aber auch einzelne ehrenamtliche Mitarbeiter, die selbst bereits älter waren. Abbildung 5.15 zeigt ein Gruppenfoto mit einem Teil der Studienteilnehmer sowie der Heimleitung und dem Masterstudenten Sergey Bogomolov (Mitte), mit dem diese Studie realisiert und durchgeführt wurde. Details zur Demographie der Teilnehmer sind im Abschnitt Ergebnisse zu finden.

Abbildung 5.15: Gruppenfoto mit einem Teil der Teilnehmer der Studie

Um den Ablauf der Studie und das Beantworten der Fragebögen für alle Teilnehmer so einfach wie möglich zu gestalten und somit bedeutungsvolle Ergebnisse erzielen zu können, wurde die Studie in enger Zusammenarbeit mit der Leitung des Stifts entworfen und durchgeführt. Zur Erleichterung der Beantwortung der Fragebögen wurden diese im Vergleich zur ersten Evaluation stark an die Bedürfnisse Senioren angepasst. Die Fragen wurden gekürzt und vereinfacht. Außerdem wurden den Senioren immer nur die Fragen vorgelegt, die sie in der jeweiligen Situation benötigten. Durch diese Aufteilung des Gesamtfragebogens sollten Irritationen durch unnötige Information und Fragen vermieden werden.

Abbildung 5.16 zeigt den experimentellen Aufbau der Studie. Sie fand in einem Aufenthaltsraum des Altersstifts statt. Reeti wurde direkt vor den Teilnehmern platziert. Der Studienleiter und das zur Steuerung des Roboters benötigte Laptop waren etwas abseits hinter dem Roboter positioniert. Auf diese Weise sollte eine möglichst ungestörte Interaktion zwischen Teilnehmer und Roboter ermöglicht werden.

Aus dem selben Grund stellte der Studienleiter nur zu Beginn der Studie kurz das Thema der Studie und den Roboter vor und hielt sich ansonsten weitestgehend im Hintergrund. Nachdem die Senioren einige einleitende, demographische Fragen beantwortet hatten, übernahm Reeti die Initiative und stellte sich nochmals selbst vor.

Abbildung 5.16: Aufbau der Studie

Zusätzlich versuchte er zusammen mit den Senioren, eine angemessene Lautstärke und Geschwindigkeit für seine Äußerungen zu finden. So sollte sichergestellt werden, dass die Empfehlungen anschließend gut verstanden werden. Außerdem sollte durch diese Interaktion bereits das Eis zwischen den Teilnehmern und Reeti gebrochen werden. Nachdem der Roboter noch einmal den Studienablauf erklärt hatte, präsentierte er die Empfehlungen jeweils in ihren acht verschiedenen Formulierungen. Nach jeder Variante wurde der Dialog pausiert, damit die Senioren in Ruhe ihre Bewertungen für die Höflichkeit und Überzeugungskraft der Formulierung abgeben konnten (wiederum auf einer 7er-Likert-Skala). Bei Bedarf konnte die Formulierung wiederholt werden. Um eventuelle Einflüsse der Präferenzen der Teilnehmer für die drei Empfehlungen besser einschätzen zu können, wurden die Senioren für jede Empfehlung gebeten, mittels einer zusätzlichen Likert-Skala zu bewerten, wie ihnen die empfohlene Maßnahme unabhängig von der Formulierung gefallen hat. Zur Vermeidung von Reihenfolgeeffekten wurde die Reihenfolge der Empfehlungen zwischen den Nutzern variiert. Zum Schluss bedankte sich Reeti für die Teilnahme und verabschiedete sich.

Ergebnisse Für die Studie konnten insgesamt 14 Bewohner und ehrenamtliche Mitarbeiter des St.Jakobs Stifts in Augsburg als Teilnehmer gewonnen werden. Die Teilnehmer (elf Frauen und drei Männer) waren zwischen 51 und 100 Jahre alt. Eine genaue Aufschlüsselung der Altersverteilung ist in Abbildung 5.17 zu sehen. Alle gaben Deutsch als ihre Muttersprache an. Da die Studie pro Teilnehmer 45 bis 60 Minuten dauerte, mussten vier der Teilnehmer die Studie leider aus Ermüdung vorzeitig abbrechen. Daher konnten lediglich die Antworten und Bewertungen der verbleibenden zehn Teilnehmer für die Auswertung der Studie berücksichtigt werden.

Wie anhand der in Abbildung 5.18 dargestellten durchschnittlichen Bewertungen zu sehen ist, vergaben die Senioren für die wahrgenommene Höflichkeit und für die wahrgenommene Überzeugungskraft durchwegs hohe Bewertungen. Ähnlich zu den Ergebnissen der ersten Evaluation wurden Fragen als am höflichsten wahrgenommen. Auch Formulierungen als Ziele des Systems erhielten wieder hohe Bewertungen hinsichtlich ihrer Höflichkeit. Allerdings gab es auch Strategien, deren mittlere Bewertung der Höflichkeit sich im Vergleich zur ersten Evaluation enorm verbessern

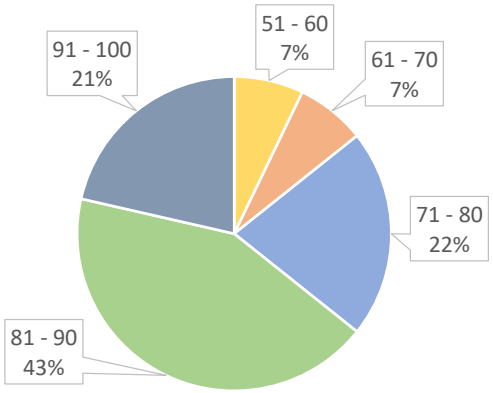

Abbildung 5.17: Altersverteilung der Teilnehmer der Studie

konnten. So zum Beispiel sokratische Hinweise ($3,58 \Rightarrow 5,58$) und direkte Kommandos ($2,27 \Rightarrow 4,58$). Direkte Kommandos galten aber dennoch weiterhin als am wenigsten höflich.

Hinsichtlich der Überzeugungskraft ließen sich anhand der deskriptiven Statistik kaum Unterschiede ausmachen. Alle Höflichkeitsstrategien erreichten durchschnittliche Bewertungen zwischen 4,73 (Andeutungen bzgl. der Absichten des Nutzers) und 5,16 (Indirekte und sokratische Andeutungen).

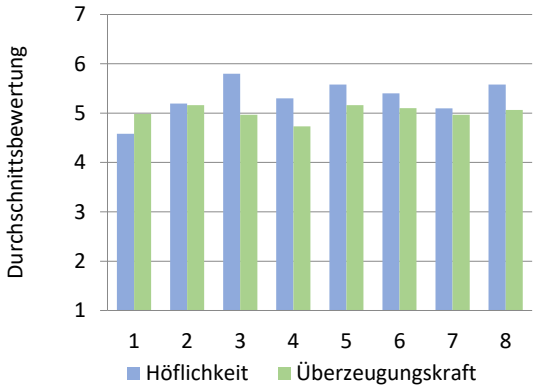

Abbildung 5.18: Durchschnittliche Bewertung der Höflichkeit und Überzeugungskraft der Formulierungen. 1: Direkte Kommandos, 2: Indirekte Andeutungen, 3: Fragen, 4: Andeutungen bzgl. der Absichten des Nutzers, 5: Sokratischer Hinweis, 6: Anfrage/Bitte, 7: Formulierungen als gemeinsames Ziel, 8: Formulierungen als Ziel des Systems

Ein ANOVA-Test für Messwiederholungen zeigte signifikante Unterschiede für die wahrgenommene Höflichkeit ($F(7,203) = 4,69, p < 0,0001$) auf. Der dazugehörige Bonferroni post-hoc Test ergab, dass Fragen und Formulierungen als Ziele des Systems signifikant höflicher wahrgenommen wurden als direkte Kommandos, siehe auch Tabelle 5.10. Hinsichtlich der wahrgenommenen Überzeugungskraft konnten durch den ANOVA-Test keine signifikanten Unterschiede gefunden werden ($F(4,39,127,38) = 0,53, p = 0,727$).

Tabelle 5.10: Wahrnehmung der Höflichkeitsstrategien (Abkürzungen: M = Mittelwert; SA = Standardabweichung; Sig = signifikant besser als)

	Höflichkeit			Überzeugungskraft		
	M	SA	Sig	M	SA	Sig
1. Direkte Kommandos	4,58	1,37		4,98	1,39	
2. Indirekte Andeutungen	5,19	1,04		5,16	1,13	
3. Fragen	5,80	0,96	1**	4,97	1,38	
4. Andeutungen bzgl. der Absichten des Nutzers	5,30	1,47		4,73	1,66	
5. Sokratische Hinweise	5,58	1,28		5,16	1,25	
6. Anfragen/Bitten	5,40	1,28		5,10	1,21	
7. Formulierungen als gemeinsames Ziel	5,10	1,38		4,97	1,58	
8. Formulierungen als Ziel des Systems	5,58	0,96	1**	5,06	1,18	
*signifikant mit p <0.05; **signifikant mit p <0.01; ***signifikant mit p <0.001						

5.2.4 Diskussion

Im Vergleich zur textbasierten Evaluation mit jüngeren Nutzern ergaben sich bei der Wizard-of-Oz-Studie mit älteren Menschen insgesamt kleinere und auch weniger signifikante Unterschiede für die wahrgenommene Höflichkeit der verschiedenen Höflichkeitsstrategien. Nichtsdestotrotz bestätigte sich die Hypothese, dass die verschiedenen Strategien die wahrgenommene Höflichkeit einer Empfehlung beeinflussen können. Wie in der ersten Evaluation wurden Fragen und Formulierungen als Ziele des Systems als am höflichsten und direkte Kommandos als am wenigsten höflich wahrgenommen. Hinsichtlich der wahrgenommenen Überzeugungskraft ergaben sich dagegen keinerlei signifikante Unterschiede zwischen den Strategien, so dass in dieser Evaluation die entsprechende Hypothese widerlegt wurde. Dennoch zeigte sich auch in dieser Evaluation, dass weniger höfliche Strategien wie direkte Kommandos dennoch als überzeugend wahrgenommen werden können. Die in Kapitel 5.2.2 vorgeschlagenen Strategien zur Auswahl von Höflichkeitsstrategien können also auch hier als vielversprechend angesehen werden.

Auf der Suche nach Erklärungen für die engeren Ergebnisse und die durchgängig besseren Bewertungen in der zweiten Evaluation wurden die gesammelten Daten und vor allem die Beobachtungen während der Studie nochmals genauer analysiert.

Zunächst kam zu Tage, dass die Senioren eine unterschiedliche Auffassung im Bezug auf die Verbindung zwischen Höflichkeit und Überzeugungskraft hatten. Während einige Teilnehmer Strategien, die ihrer Meinung nach höflich waren, auch als überzeugend einstuften, werteten andere Teilnehmer genau entgegengesetzt.

Ein weiterer Grund für die geringen Unterschiede zwischen den Bewertungen der Höflichkeitsstrategien könnten Beeinflussungen durch die empfohlenen Aktionen

selbst sein. Zwei Senioren waren aufgrund körperlicher Einschränkungen nicht mehr oder nur schwer dazu in der Lage, ein Fenster alleine zu öffnen oder einen größeren Spaziergang zu machen. Dadurch fiel es ihnen dann auch schwer, die Formulierungen für diese Empfehlungen unabhängig von ihren eigenen Einschränkungen zu bewerten und zu differenzieren. Auch Schwierigkeiten beim Hören und eine fortschreitende Ermüdung während der Studie führten bei einem Teil der Teilnehmer zu einem ähnlichen Effekt.

Einen ebenfalls nicht zu vernachlässigender Einfluss könnte durch den Einsatz des Roboters verursacht worden sein. Nomura und Kollegen [Nomura und Takeuchi, 2011] fanden in einer Reihe von Experimenten heraus, dass ältere Nutzer häufig positivere Eindrücke von Robotern haben als jüngere Nutzer. Diese hauptsächlich positiven Erfahrungen konnten auch bei unseren Senioren festgestellt werden. Alle Teilnehmer der Studie waren weder verängstigt, noch lehnten sie den Roboter ab. Da die meisten von ihnen vorher noch nie einen ähnlichen Roboter gesehen hatten, gab es zwar zu Beginn etwas Skepsis und Zweifel, dass man sich „mit so einer Maschine unterhalten kann“. Nach wenigen Minuten unterhielten sich viele der Senioren allerdings mit Reeti, als würde er sie verstehen. Sie erzählten teilweise sogar private Geschichten über ihre Kindheit, ihr aktuelles Leben und ihre Familien. Ausgelöst wurde dieses Verhalten teilweise durch die zufälligen Bewegungen Reetis, die in manchen Situationen wie non-verbale Reaktionen auf das Gesagte der Teilnehmer wahrgenommen wurden. Viele der Teilnehmer lobten außerdem, dass Reeti „sehr freundlich“, „sehr nett“ und „wirklich höflich“ sei. Dieses Lob wurden teilweise sogar direkt an Reeti gerichtet. Insgesamt stellte die Zeit und die Interaktion mit dem Roboter für viele der Teilnehmer eine schöne und willkommene Erfahrung und Abwechslung in ihrem Alltag dar. Es stellt sich allerdings die Frage, ob diese positive Wahrnehmung des Roboters langfristig anhalten würde und wie sich eine tägliche Interaktion mit ihm auf die Nutzung und Wahrnehmung des Systems auswirken würde.

5.3 Persönlichkeitsausprägungen von Formulierungen

Die Persönlichkeit eines Menschen ist die Gesamtheit seiner individuellen charakteristischen Eigenschaften. Diese sind zum Teil vererbt, zum Teil aber auch durch die Kultur und persönliche Erfahrungen geprägt [Hofstede, 2001].

Für die HCI ist die Persönlichkeit bedeutsam, da sie dauerhaft das Verhalten sowie den Geschmack und die Interessen eines Menschen beeinflusst [Jung und Baynes, 1923]. Wie in Kapitel 4.1 beschrieben, wurde die Persönlichkeit der Nutzer deshalb auch schon in Ansätzen zur Empfehlungsauswahl berücksichtigt [Dunn et al., 2009, Hu und Pu, 2011, Lin und McLeod, 2002, Nunes, 2008].

Für beratende Empfehlungssysteme ist die Persönlichkeit aber auch aus einem weiteren Grund interessant. Laut Nass und Lee [Nass und Lee, 2001] hat die Persönlichkeit auch einen Einfluss darauf, wie Nutzer sprach- und textbasierte Systeme wahrnehmen und wie sie anschließend mit ihnen interagieren. Personen fühlen sich zum Beispiel angezogen, wenn das System eine ähnliche Persönlichkeit aufweist wie sie selbst. In Forschungsarbeiten zu virtuellen Agenten oder sozialen Robotern wurden diese Erkenntnisse bereits berücksichtigt, um das verbale und non-verbale Verhalten der Agenten gezielt an die Persönlichkeit der Nutzer anzupassen, siehe zum Beispiel [Bee et al., 2010, Krenn et al., 2014].

Hinsichtlich der Anpassung von Empfehlungstexten konnten bei der Recherche für diese Arbeit noch keine Arbeiten gefunden werden. Die Erforschung von Empfehlungstexten bezieht sich größtenteils auf den inhaltlichen Aufbau von Erklärungen, siehe Kapitel 5.1. Da in assistierenden Empfehlungssystemen jedoch sowohl das Nutzervertrauen gegenüber den Systemen als auch die Überzeugungskraft der Systeme eine wichtige Rolle spielen, ist anzunehmen, dass eine gezielte Anpassung der Persönlichkeitsausprägung des Systems an die jeweiligen Nutzer auch für Systeme wie CARE oder SavER von Vorteil sein kann. Aus diesem Grund wird in diesem Kapitel der bisher im Bereich Empfehlungssysteme noch vernachlässigte Aspekt der Adaption von Empfehlungstexten basierend auf Persönlichkeitsmerkmalen erforscht.

Forschungsfrage Im Detail wird in dieser Dissertation untersucht, wie eine Adaption der Systempersönlichkeit technisch ermöglicht werden kann und welche Wirkung spezifische Anpassungsstrategien auf die wahrgenommene Vertrauenswürdigkeit und die wahrgenommene Überzeugungskraft einer Empfehlung haben. Im besonderen wird der Frage nachgegangen, ob die beiden Kriterien durch eine Spiegelung der Persönlichkeit der Nutzer verbessert werden können.

Im Folgenden wird zunächst das Big-Five-Persönlichkeitsmodell beschrieben. Anschließend werden theoretische Grundlagen hinsichtlich der Adaption der Persönlichkeit von Systemen vermittelt. In Kapitel 5.3.3 wird dann eine prototypische Umsetzung einer persönlichkeitsbasierten Textgenerierung für Empfehlungen erklärt. Dieser Prototype wurde innerhalb einer Evaluation eingesetzt, um die Wirkung verschiedener Adaptionsstrategien zu untersuchen. Das Design, die Durchführung und

die Ergebnisse dieser Evaluation werden in Kapitel 5.3.4 beschrieben. Eine Diskussion der Erkenntnisse in Kapitel 5.3.5 schließt dieses Kapitel ab.

5.3.1 Big-Five-Persönlichkeitsmodell

Eine einfache und vollständige Beschreibung der Persönlichkeit eines Menschen zu erstellen ist nahezu unmöglich. Im Laufe der Zeit wurden verschiedenste Ansätze zum Verständnis der menschlichen Persönlichkeitsstruktur entwickelt [Hutterer, 2013, Pervin et al., 2005, Schmidt und Birbaumer, 1996]. Einen wichtigen Ansatz für die HCI stellen auf Persönlichkeitsmerkmalen [Eysenck und Eysenck, 1965] basierende Faktorenmodelle dar [Eysenck und Eysenck, 1965, Gosling et al., 2003, Lee und Ashton, 2004]. Sie beschreiben die Persönlichkeit von Menschen anhand unterschiedlicher Persönlichkeitsmerkmale wie der *Introversion/Extraversion* oder der *Verträglichkeit*. Der Vorteil von Faktorenmodellen besteht darin, dass sich Persönlichkeiten durch Werte für die Ausprägung jedes Persönlichkeitsmerkmals darstellen lassen. Dadurch lassen sie sich gut in Studien nutzen und in Systeme integrieren.

Eines der am weitesten verbreiteten Modelle ist das Big-Five-Persönlichkeitsmodell [86], auch Fünf-Faktorenmodell genannt, [Gosling et al., 2003]. Es beschreibt die Persönlichkeit einer Person anhand der Faktoren *Offenheit*, *Gewissenhaftigkeit*, *Verträglichkeit*, *Neurotizismus* und *Extraversion*, siehe Abbildung 5.19. Diese Faktoren haben sich mit der Zeit als die wesentlichen Faktoren der menschlichen Persönlichkeit herausgestellt [Asendorpf und Neyer, 2012]. Zwar ergeben sich zum Beispiel aufgrund der Herkunft und kultureller Unterschiede auch Unterschiede für die Ausprägung der Faktoren in der Persönlichkeit. Allerdings konnten u.a. Lang und Lüdtke bestätigen, dass sich charakteristische und wiederkehrende Verhaltensweisen klar durch eindeutige Bewertungen für die fünf Faktoren beschreiben lassen [Lang und Lüdtke, 2005].

Offenheit Die Offenheit von Menschen zeigt sich in ihrer Aufgeschlossenheit gegenüber neuen Erfahrungen und Erlebnissen. Für offene Menschen sind ein großer Einfalls- und Erfindungsreichtum sowie Neugier und Kreativität charakteristisch [Mairesse, 2008]. Sie weichen auch häufiger von der Norm ab [Costa und MacCrae, 1992]. Extreme Ausprägungen dieses Persönlichkeitsmerkmals sind allerdings eher selten [McCrae, 1996].

Gewissenhaftigkeit Personen, die sie sich durch Zielstrebigkeit, Sorgfalt und Zuverlässigkeit auszeichnen, werden als gewissenhaft wahrgenommen. Sie sind dazu in der Lage impulsives Verhalten zu verhindern und gelten deswegen häufig als vorsichtig, selbst-diszipliniert und erfolgsgesteuert. [Mairesse, 2008]

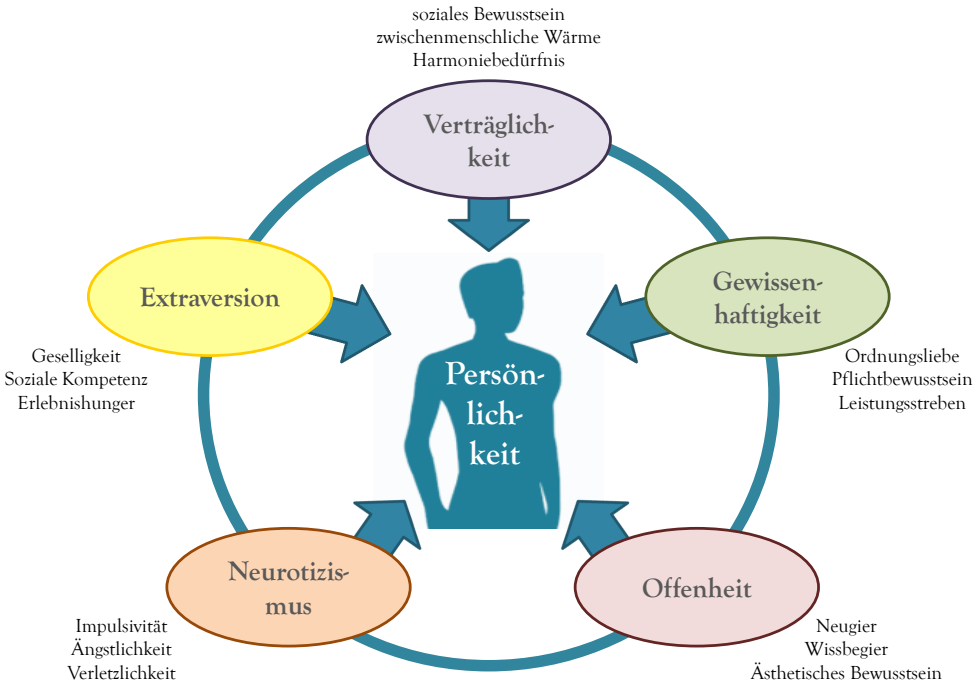

Abbildung 5.19: Faktoren des Big-Five-Persönlichkeitsmodells und beispielhafte charakteristische Merkmale für jeden Faktor (nach [Gosling et al., 2003])

Verträglichkeit Ein hohes Maß an Verträglichkeit äußert sich dadurch, dass Personen anderen Menschen leicht ihr Vertrauen schenken und immer das Beste in ihnen sehen. Verträgliche Menschen sind nicht nachtragend und meistens optimistisch. Dazu passend sind Wohlwollen, Mitgefühl und Nachgiebigkeit typische Attribute verträglicher Menschen. Ihr Interesse gilt ihren Mitmenschen und deren Wohlbefinden, was sich auch durch eine ausgeprägte Großzügigkeit zeigt. Dies führt häufig dazu, dass verträgliche Personen auch als sehr angenehm wahrgenommen werden.

Neurotizismus Neurotische Menschen sind häufig ängstlich, besorgt oder traurig und fühlen sich schnell einsam. Des Weiteren besitzen sie eine geringe Frustrationsgrenze und können nur schlecht mit Stress umgehen. Damit einher gehen eine Neigung zu Gefühlsausbrüchen, eine erhöhte Anfälligkeit für Krankheiten wie Burnout und letztendlich eine geringe Lebenszufriedenheit. [Asendorpf, 2015, Costa und MacCrae, 1992]

Extraversion Das wichtigste Persönlichkeitsmerkmal der „Big-Five“ ist die Extraversion. Sie ist auch Teil der meisten anderen Persönlichkeitsmodelle und wird in den meisten Persönlichkeitstests verwendet. Nach der Theorie von Eysenck tragen Menschen sowohl extrovertierte als auch introvertierte Züge in sich. Allerdings gibt es je nach Persönlichkeit starke Neigungen zu einem der beiden Extrema [Eysenck und Eysenck, 1965]. Das Wort „Extraversion“ beschreibt bereits die „Nach-Außen-Gewandtheit“ extrovertierter Menschen. Sie teilen sich gerne

ihren Mitmenschen mit und lieben den Austausch mit ihnen. Sie suchen die Nähe größerer Gruppen und sind aktiv, durchsetzungsfähig und abenteuerlustig [Costa und MacCrae, 1992]. Darin liegt auch der Grund für die Wichtigkeit der Extraversion als Persönlichkeitsmerkmal. Eine stark ausgeprägte Extraversion kann kaum übersehen werden. Außerdem wird sie bis zu einem gewissen Maß als positiv wahrgenommen, da mit ihr u.a. Stärke und Attraktivität assoziiert werden. Das macht sie auch für Marketingstrategien zu einem vielversprechenden Mittel zur Gewinnung der Kunden [Aaker, 1997].

In Tabelle 5.11 ist eine Auswahl an Adjektiven aufgelistet, die spezifischen Ausprägungen der einzelnen Persönlichkeitsmerkmale zugeordnet werden können.

Tabelle 5.11: Eigenschaften der Ausprägungen der Big Five-Persönlichkeitsmerkmale

Persönlichkeits -merkmal	**Eigenschaften (hohe Ausprägung)**	**Eigenschaften (geringe Ausprägung)**
Extraversion	gesprächig, freimütig, unternehmungslustig, gesellig	schweigsam, verschlossen, zurückhaltend, zurückgezogen
Verträglichkeit	gutmütig, wohlwollend, freundlich, kooperativ	grantig, missgünstig, starrköpfig, feindselig
Gewissenhaftigkeit	sorgfältig, zuverlässig, genau, beharrlich	nachlässig, unzuverlässig, ungenau, sprunghaft
Neurotizismus	nervös, ängstlich, erregbar, wehleidig	ausgeglichen, entspannt, gelassen, körperlich robust
Offenheit	kunstverständig, intellektuell, kultiviert, phantasievoll	kunstunverständig, ungebildet, ungeschliffen, phantasielos

5.3.2 Adaption der Persönlichkeit von Systemen

Die Adaption der Persönlichkeitsmerkmale innerhalb der Sprach- und Textausgabe eines Systems besteht aus mehreren Schritten. Zunächst schätzt das System die Persönlichkeit der Nutzer ein. Dann wählt es eine geeignete Adaptionsstrategie aus und schließlich passt es seine Sprach- oder Textausgabe so an, dass die erwünschten Ausprägungen der Persönlichkeitsmerkmale erreicht werden.

Ermittlung der Persönlichkeit Eine präzise Einschätzung der Persönlichkeit eines Menschen ist nur durch eine langwierige Beobachtung der jeweiligen Person in verschiedensten Situationen möglich. Dabei muss sowohl das verbale als auch das non-verbale Verhalten (Gestik, Mimik oder Blickverhalten) beobachtet und analysiert werden. Da eine langwierige Persönlichkeitsanalyse für viele Systeme jedoch nicht praktikabel ist, wurden verschiedene Modelle zur Erkennung von Persönlichkeiten aus geschriebener Sprache [Pennebaker und King, 1999] und Konversationen [Mairesse et al., 2007] entwickelt. Ein gängiges Vorgehen ist der Vergleich der Aussagen einer Person mit Beispielen aus Korpora, in denen bereits Signalwörter und

Satzmuster anhand ihrer sprachlichen Merkmale mit Ausprägungen der Persönlichkeitsmerkmale in Verbindung gebracht wurden [Mairesse et al., 2007].

Eine weitere und einfachere Methode zur Einschätzung der Persönlichkeit sind Persönlichkeitstests. Eysenck [Eysenck und Eysenck, 1965] entwickelte bereits für seine einfachen Faktorenmodelle mit den Dimensionen *Extraversion-Introversion* und *Labilität-Stabilität (Neurotizismus)* Fragebögen, mit denen er die Ausprägung dieser Faktoren messen konnte. Unter dem Begriff der *Personality Inventories* [Asendorpf und Neyer, 2012] wurden mit der Zeit neben den Persönlichkeitsmodellen auch die dazugehörigen Fragebögen weiterentwickelt und fanden mehr und mehr Anwendung in der Praxis. Der Vorteil dieser Fragebögen ist, dass ihre Nutzung sehr einfach und schnell von statten geht.

Adaptionsstrategien Assistierende Empfehlungssysteme weisen Merkmale verschiedener Anwendungsfälle für die Adaption der Persönlichkeit auf, für die es jeweils eigene Adaptionsstrategien gibt, siehe Tabelle 5.12.

Zum einen sollen den Nutzern hilfreiche Informationen präsentiert werden, um ihre Kompetenz in der Anwendungsdomäne zu verbessern und die Entscheidungen für und gegen Empfehlungen zu erleichtern. Je nach Erfahrungsgrad der Nutzer sollte das System daher entweder extrovertiert, aber verträglich oder an die Persönlichkeit der Nutzer angepasst agieren [Mairesse und Walker, 2010]. Die Hauptaufgabe der Systeme ist jedoch, die Nutzer von der Notwendigkeit der empfohlenen Maßnahmen und Aktivitäten zu überzeugen. Eine große Überzeugungskraft und Motivation kann u.a. durch soziale Dominanz erreicht werden [Bee et al., 2010]. Diese soziale Dominanz kann wiederum entweder durch Extraversion oder eine geringe Verträglichkeit hervorgerufen werden [Furnham, 1990]. Im Hinblick auf eine langfristig gute Beziehung zwischen System und Nutzer sollte ein beratendes Empfehlungssystem aber dennoch nicht als Lehrer oder Verkäufer wahrgenommen werden. Es sollte so auftreten, dass es als verträglicher und pflichtbewusster Assistent akzeptiert wird. Dies könnte u.a. dadurch erreicht werden, dass es situativ auf den Gemütszustand der Nutzer eingeht. In CARE ist zum Beispiel Vorsicht geboten, da sich der psychologische Zustand (z.B. Fröhlichkeit und Aktivität oder Niedergeschlagenheit und Lethargie) der Senioren von Tag zu Tag unterscheiden kann. In diesen Fällen kann zwar dieselbe Empfehlung sinnvoll sein. Ein zu forsches oder dominantes Verhalten, könnte allerdings sowohl der Motivation der Nutzer als auch der Beziehung zwischen System und Nutzer schaden.

Persönlichkeit und ihre Auswirkungen auf die Sprache Den Zusammenhang zwischen Sprache und Persönlichkeit stellt zum Beispiel die Sedimentationshypothese bzw. lexikalische Hypothese her [Galton, 1949]. Laut ihr schlagen sich Persönlichkeitsmerkmale in der Sprache nieder und können teilweise alleine basierend auf den verwendeten Wörtern gedeutet werden. Im Folgenden werden die Auswirkungen der einzelnen Faktoren des Big-Five-Persönlichkeitsmodells auf die Sprache [Mairesse, 2008, Mairesse und Walker, 2007] beschrieben.

Tabelle 5.12: Kontextbasierte Strategien zur Adaption der Persönlichkeit eines Systems (nach [Mairesse und Walker, 2010])

Anwendungsfall	**Persönlichkeit (Nutzer)**	**Persönlichkeit (System)**
Präsentation von Informationen	Anfänger/Neuling	extrovertiert, verträglich
	Fortgeschrittener/Experte	dem Nutzer angepasst
Lehrsysteme	jede	extrovertiert, verträglich, pflichtbewusst
Tele-Verkauf-System	jede	extrovertiert, der Marke entsprechend
Abfrage wichtiger Daten	jede	pflichtbewusst, nicht extrovertiert

Offenheit Für die Offenheit konnten bisher nur wenige Einflüsse auf sprachlicher Ebene nachgewiesen werden. Offene Menschen neigen zur Verwendung positiver Wörter, nutzen längere Äußerungen und zeigen häufiger Empathie gegenüber ihren Gesprächspartnern, ob bekannt oder unbekannt.

Gewissenhaftigkeit Gewissenhafte Menschen verwenden häufig positive Formulierungen und kommen weitestgehend ohne wiederholte oder paraphrasierte Formulierungen aus. Ihre Sorgfalt spiegelt sich auch in Nachfragen wie „Hast du gerade X gesagt?“ zu Beginn einer eigenen Aussage wider.

Verträglichkeit In der Sprache zeigt sich eine hohe Verträglichkeit durch die häufige Verwendung positiver Wörter und einen generell eher positiven Inhalt der Aussagen. Diese positiven Aussagen werden auch meist zuerst geäußert. Das Interesse an anderen Menschen zeigen verträgliche Menschen durch regelmäßiges Fragen nach Bestätigung und die Einbeziehung der Gesprächspartner in das Gespräch.

Neurotizismus Für die Planung von Dialogen bzw. für die Generierung von Empfehlungstexten bedeutet eine neurotische Persönlichkeit vor allem eine stärkere Negativität in der Sprache. Positive Inhalte werden genauso wie Skepsis ausdrückende Aussagen wie „Ich bin mir nicht sicher...“ aus Unsicherheit und zur Wahrung der Kontrolle meist an den Anfang eines Satzes gestellt. Außerdem haben neurotische Menschen häufig das Verlangen nach Bestätigung durch den Gegenüber und erfragen diese auch durch Phrasen wie „Findest du nicht auch?“. Zusätzlich treten häufiger lexikalische Wiederholungen auf.

Extraversion Da sich die Extraversion von allen Persönlichkeitsmerkmalen am deutlichsten in der Sprache niederschlägt, ist sie auch für die Verwendung in der HCI sehr attraktiv. Eine extrovertierte Persönlichkeit fällt vor allem durch lautes und wortreiches Reden auf. Außerdem wechseln extrovertierte Personen innerhalb eines Dialogs häufiger das Thema als weniger extrovertierte Personen. Sie

wiederholen sich aber auch häufiger. Außerdem verwenden sie vermehrt positive Wörter und beginnen Aussagen häufig mit positiven Inhalten. Eine verstärkte Extraversion ist aber auch mit Mitgefühl verbunden, was sich häufig dadurch zeigt, dass eine Bestätigung von Aussagen der Gesprächspartner eingeholt wird.

5.3.3 Prototypische Umsetzung

Für den Vergleich verschiedener Strategien zur Anpassung der Persönlichkeit von Empfehlungstexten wurde ein Prototyp entwickelt, der anschließend an die Ermittlung der Persönlichkeit der Nutzer Formulierungen für Empfehlungen generieren konnte, die entweder einer neutralen Persönlichkeit, der Persönlichkeit der Nutzer oder der entgegengesetzten Persönlichkeit der Nutzer entsprachen.

Der Ablauf der Generierung eines Empfehlungstextes ist in Abbildung 5.20 dargestellt. Als Eingabewerte erhält der Prototyp die Ergebnisse eines Persönlichkeitstest (Ausprägungen der Faktoren des Big-Five-Persönlichkeitsmodells von 1 bis 7). Basierend auf diesen Werten generiert das Framework PERSONAGE [Mairesse und Walker, 2007] gezielt formulierte Empfehlungstexte, die über den VisualSceneMaker [Gebhard et al., 2012, Mehlmann et al., 2015] an den sozialen Roboter Reeti weitergeleitet werden. Dieser gibt die generierten Aussagen dann wieder. Da PERSONAGE für die Sprachgenerierung im Englischen entwickelt wurde, waren auch die untersuchten Empfehlungstexte auf Englisch.

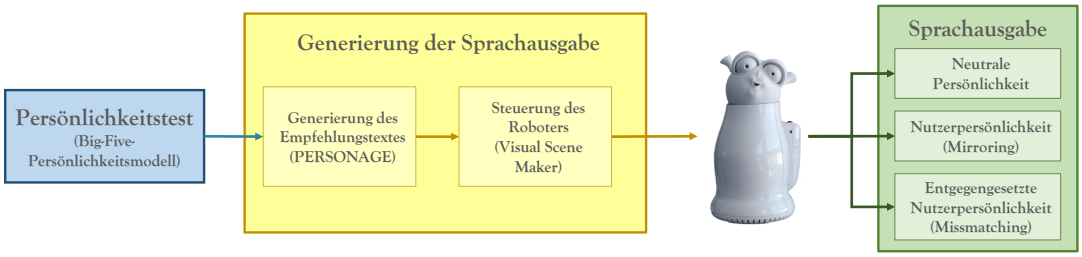

Abbildung 5.20: Ablauf und Informationsfluss der Generierung von Empfehlungstexten mit spezifischen Persönlichkeitsmerkmalen

Im Folgenden wird die Sprachgenerierung mit PERSONAGE erklärt. Für eine kurze Beschreibung des VisualSceneMaker und des Roboters Reeti sei auf Kapitel 5.2.3 verwiesen. Der Persönlichkeitstest ist in Anhang B zu finden.

PERSONAGE Das Natural-Language-Generator-Framework (NLG-Framework) PERSONAGE wurde von Mairesse und Walker entwickelt und ist ein Werkzeug, das basierend auf der Kombination von Persönlichkeitsmerkmalen verschiedene Formulierungen einer Aussage generieren kann [Mairesse und Walker, 2007]. Das besondere an PERSONAGE ist, dass es durch eine Vielzahl von Eingabeparametern eine sehr feingranulare Formulierung von Sprachausgaben erlaubt, die sich in der Wahrnehmung der Persönlichkeitsmerkmale durch die Nutzer niederschlagen können. Die einzelnen Module und Arbeitsschritte zur Generierung einer natürlichen

Sprachausgabe zeigt Abbildung 5.21. Für ein grobes Verständnis des Frameworks werden die einzelnen Module im Anschluss kurz vorgestellt. Für weitere Details sei auf Mairesses Dissertation [Mairesse, 2008] und weitere Veröffentlichungen verwiesen [Mairesse und Walker, 2007, Mairesse und Walker, 2010].

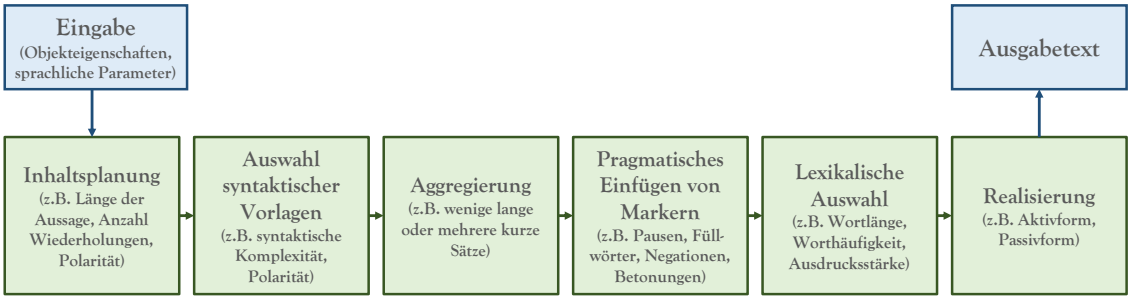

Abbildung 5.21: Architektur des NLG-Frameworks PERSONAGE (nach [Mairesse und Walker, 2007])

Inhaltsplanung Mit der Planung des Inhalts wird die Grundlage für die spätere Ausformulierung erstellt. Zunächst wird aus den verschiedenen Eingabedaten das Kommunikationsziel ermittelt. Anschließend wird ein sog. *Content Plan Tree* erstellt, der die Struktur und den Inhalt der einzelnen Teile der Aussage beinhaltet. Das Beispiel in Abbildung 5.22 (links) zeigt den Inhalt einer Empfehlung für einen Spaziergang inklusive zweier Argumente als Rechtfertigung.

Bereits bei der Inhaltsplanung kann die spätere Persönlichkeit der Aussage beeinflusst werden. So kann zum Beispiel die Länge der Aussage, die Anzahl der Wiederholungen und Paraphrasierungen, aber auch die Polarität der Inhalte (negativ oder positiv) sowie die Positionierung und Gruppierung dieser Inhalte gesteuert werden.

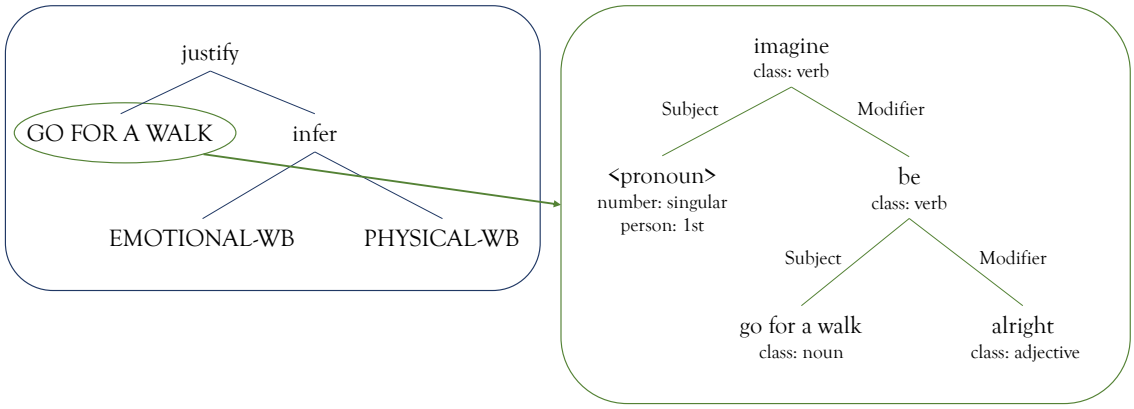

Abbildung 5.22: Beispiel eines Content Plan Tree (links) und einer Deep Syntactic Structure (rechts) der Aussage „I imagine going for a walk is alright. It offers satisfying emotional well-being with excellent physical well-being“ (nach [Mairesse und Walker, 2007])

Auswahl syntaktischer Vorlagen Als zweiter Schritt folgt die Gestaltung der syntaktischen Struktur der Satzbausteine. Hierfür werden die Satzteile mit syntaktischen Vorlagen, sog. *Deep Syntactic Structures (DSyntS)* assoziiert, die zum Erreichen der Zielaussage am geeignetsten sind, siehe Abbildung 5.22 (rechts). Die Bausteine der DSyntS werden anschließend mit Lexemen, d.h. der Grundform eines Wortes, befüllt und daraus erste String-Repräsentationen der Satzteile generiert.

Auch in diesem Arbeitsschritt kann wieder Einfluss auf die letztendliche Persönlichkeitsausprägung der Zielaussage genommen werden. Durch die Auswahl geeigneter Vorlagen kann zum Beispiel eine stärkere Selbstreferenzierung oder eine komplexere Syntax erreicht werden. Aber auch die Polarität der Aussage kann wiederum beeinflusst werden.

Aggregierung In dieser Phase können einzelne Satzbausteine oder Sätze durch Verbindungsstücke verbunden oder zusammengefasst werden.

Je nach erwünschter Persönlichkeitsausprägung können in dieser Phase entweder weniger längere oder mehrere kürzere Sätze erzeugt werden.

Pragmatisches Einfügen von Markern Als Marker werden kleine Ausdrücke bezeichnet, die die Wirkung einzelner Satzbausteine und Persönlichkeitsmerkmale verstärken oder abschwächen können. Dazu gehören zum Beispiel bewusste Pausen, Negationen oder Füllwörter wie „äh“ oder „ich meine“. Auch Ausdrücke, die die Sicherheit eines Sprechers bzgl. seiner Aussage verdeutlichen, können nach der Aggregation integriert werden. Dazu zählen Abschwächungen wie „eine Art“ oder „relativ“ und Betonungen wie „wirklich“, „grundsätzlich“ oder „Jeder weiß, dass...“.

Lexikalische Auswahl In dieser Phase werden die letztendlich verwendeten Wörter selektiert. Dazu werden Synonyme hinsichtlich der Anforderungen an das Wort miteinander verglichen. Diese Anforderungen können zum Beispiel die Länge, die Häufigkeit der Verwendung oder die Ausdrucksstärke des Worts betreffen.

Für eine introvertierte Begründung für einen Spaziergang könnten zum Beispiel exklusivere und längere Wörter wie „fördert ihr körperliches Wohlbefinden“ oder „löst emotionale Anspannungen“ eingesetzt werden, während für einen extrovertierten Ausdruck kürzere und häufig verwendete, aber emotional ausdrucksstarke Wörter wie „hält Sie fit“ oder „wirkt entspannend“ geeigneter wären.

Realisierung In der letzten Phase wird aus jedem der fertigen DSyntS unter Berücksichtigung grammatikalischer Regeln ein vollständiger Satz gebildet.

Als eine der letzten Variationsmöglichkeiten steht hier zum Beispiel noch die Wahl zwischen einer Passiv- oder Aktivform des Satzes zur Verfügung. Ein Beispiel wäre „Ein Spaziergang fördert ihr körperliches Wohlbefinden.“ im Vergleich zu „Durch einen Spaziergang wird ihr körperliches Wohlbefinden gefördert.“

Für das Mapping der gewünschten Persönlichkeitsausprägung auf die Eingabeparameter in PERSONAGE gibt es verschiedene Ansätze, die auf der Arbeit von Langkilde-Geary und Kollegen beruhen [Langkilde und Knight, 1998, Langkilde-Geary, 2002]:

Regelbasierte Generierung Die einfachste Variante des Mappings ist die regelbasierte Auswahl der Eingabeparameter basierend auf der gewünschten Persönlichkeit. Allerdings führen bei dieser Methode identische Eingabeparameter auch zu identischen Ausgaben. Mairesse und Walker versuchten Wiederholungen bestimmter Ausdrücke durch zufällige, leichte Abweichungen der Eingabewerte zu vermeiden. Ein weiterer Nachteil ist, dass der regelbasierte Ansatz unabhängig von der tatsächlichen Ausprägung der Persönlichkeitsmerkmale nur die Extrema der Merkmale, also zum Beispiel introvertiert und extrovertiert, berücksichtigt. Dadurch sind nur einige wenige extreme Persönlichkeitsausprägungen für Aussagen möglich.

Stochastische Generierung Kontinuierliche und damit natürlichere Persönlichkeitsausprägungen erhält man durch stochastische Ansätze wie den *Overgenerate-and-Select-Ansatz* oder die *Parameter-Estimation-Methode*. Diese setzen sich aus einer *Entwicklungsphase* und einer *Generierungsphase* zusammen.

In der Entwicklungsphase werden die folgenden Schritte durchlaufen:

1. (zufallsgesteuerte) Generierung von Aussagen, die das komplette Spektrum an Persönlichkeitsausprägungen abdecken
2. Bewertung der Äußerungen durch Nutzer anhand von Persönlichkeitstests
3. Abschätzung der Ausprägungen der Persönlichkeitsmerkmale für jede Aussage
4. Trainieren stochastischer Modelle zur Schätzung von Nutzerbewertungen

Dieses Vorgehen bringt die Ausprägungen bestimmter Persönlichkeitsmerkmale mit den Eingabeparametern der Sprachgenerierung in Zusammenhang. Dadurch werden feingranulare Vergleiche und Selektionen der Äußerungen möglich.

Die Generierungsphase unterscheidet sich dann je nach Ansatz. Beim Overgenerate-and-Select-Ansatz werden mehrere, unterschiedliche Äußerungen generiert, die die gesamte Parameterbreite ausnutzen. Für diese werden mittels der trainierten Modelle die Ausprägungen der Persönlichkeitsmerkmale ermittelt. Anschließend wird die Äußerung ausgewählt, die der gewünschten Persönlichkeit am nächsten kommt. Bei der Parameter-Estimation-Methode werden dagegen die für die einzelnen Eingabeparameter am besten geeigneten Modelle verwendet, um von der Beschreibung einer Persönlichkeit auf die Eingabeparameter für die Generierung zu schließen.

In der Evaluation in dieser Dissertation wurde die stochastische Parameter-Estimation-Methode verwendet.

5.3.4 Evaluation

Da die recherchierten Strategien zur Anpassung der Persönlichkeit von Systemen keine eindeutige Vorgabe für die Adaption der Persönlichkeit in Empfehlungstexten bieten konnten, wurde mittels einer Studie untersucht, welche Strategien am besten geeignet sind, um die Vertrauenswürdigkeit und die Überzeugungskraft der Empfehlungen zu steigern. Diese Strategien lauteten:

Neutrale Persönlichkeit In dieser Variante, die als Baseline der Studie angesehen wurde, wurde eine neutrale Persönlichkeit für die Empfehlungstexte gewählt. Es fand keine Adaption an die Nutzerpersönlichkeit statt. Da die Skalen für die Beschreibung der Faktoren des Big-Five-Persönlichkeitsmodells von 1 bis 7 reichen, wurde für die neutrale Persönlichkeit für alle Faktoren ein Wert von 4 verwendet.

Mirroring Beim Mirroring bzw. Spiegeln wird dem System eine Persönlichkeit verliehen, die der Persönlichkeit der Zielperson möglichst ähnlich ist [O'Connor und Seymour, 2011]. Hierfür werden dem System exakt die Ausprägungen der Persönlichkeitsmerkmale übergeben, die für die Zielperson durch eine Analyse der Persönlichkeit ermittelt wurden. Diese Variante der Anpassung folgt der Similarity Attraction Theory und sollte die Vertrauenswürdigkeit des Systems vor allem bei neurotischen und introvertierten Personen positiv beeinflussen können [Nass und Lee, 2001, Reeves und Nass, 1998].

Mismatching Weist eine Person oder ein System eine entgegengesetzte Persönlichkeit zu einer Person auf, spricht man von Mismatching [O'Connor und Seymour, 2011]. Für die Studie wurde Mismatching dadurch erzeugt, dass die Ausprägungen der Persönlichkeitsmerkmale der Teilnehmer am Mittelpunkt der Skalen, mit denen in PERSONAGE die Persönlichkeitsmerkmale beschrieben werden, gespiegelt wurden. Bei weniger extremen Ausprägungen fiel dadurch allerdings der Unterschied zur gegensätzlichen Ausprägung geringer aus.

Als Folge des Mismatchings ist davon auszugehen, dass eine stärkere Distanz zwischen System und Nutzer entsteht. Dadurch könnte die Vertrauenswürdigkeit des Systems weniger stark eingeschätzt werden. Allerdings zeigten Rushton und Kollegen, dass zum Beispiel introvertierte Menschen besser von extrovertierten Lehrern lernen können [Rushton et al., 1987]. Ähnlich wie bei der Höflichkeit könnte das gezielte Hervorheben von Abweichungen und das Erzeugen von Reibung für manche Menschen also womöglich die Überzeugungskraft stärken.

Hypothesen Die Annahmen, die in der Evaluation untersucht wurden, lauteten:

1. Die Art der Adaption des Empfehlungstextes an die Nutzerpersönlichkeit hat einen Effekt auf die Vertrauenswürdigkeit der Empfehlung.

2. Die Art der Adaption des Empfehlungstextes an die Nutzerpersönlichkeit hat einen Effekt auf die Überzeugungskraft der Empfehlung.

3. Das Mirroring der Nutzerpersönlichkeit in Empfehlungstexten führt, im Vergleich zu den anderen Strategien, zu einer verbesserten Vertrauenswürdigkeit.

4. Das Mismatching der Nutzerpersönlichkeit in Empfehlungstexten führt, im Vergleich zu den anderen Strategien, zu einer verbesserten Überzeugungskraft.

Durchführung Nach einer groben Einführung in den Ablauf der Studie und nach der Beantwortung demographischer Fragen (Alter, Geschlecht, Beruf) wurden die Studienteilnehmer gebeten einen Persönlichkeitstest auszufüllen. Dieser Test wurde von Satow [Satow, 2012] erstellt und beinhaltet für jeden der Faktoren des Big-Five-Persönlichkeitsmodells 10 Aussagen, die mit Hilfe einer 4er-Likert-Skala von „trifft gar nicht zu“ bis „trifft genau zu“ bewertet werden sollen. Der vollständige Fragenkatalog und eine kurze Beschreibung sind in Anhang B zu finden. Die folgende Aufzählung enthält einen Auszug aus diesem Fragenkatalog.

- Ich bin ein ängstlicher Typ. (Neurotizismus)
- Ich bin gerne mit anderen Menschen zusammen. (Extraversion)
- Ich gehe immer planvoll vor. (Gewissenhaftigkeit)
- Ich lerne immer wieder gerne neue Dinge. (Offenheit)
- Ich bin ein höflicher Mensch. (Verträglichkeit)

Basierend auf den Ergebnissen des Persönlichkeitstest wurden für alle Teilnehmer für jede der Adaptionsstrategien für drei Empfehlungen aus dem CARE-Szenario Empfehlungstexte generiert. Diese Empfehlungen waren: „Rufe jemanden an.“, „Mache ein Kreuzworträtsel.“ und „Gehe spazieren.“

Beispiele für Formulierungen für eine extrovertierte und offene Person mit einem sehr geringen Hang zum Neurotizismus und neutralen Bewertungen für Gewissenhaftigkeit und Verträglichkeit waren:

- Neutral: „Mhm...basically, it offers satisfying emotional well-being and it provides excellent physical well-being, alright? I imagine going for a walk is alright.“
- Mirroring: „Mhm...basically, I imagine going for a walk is alright. It provides satisfying emotional well-being and it offers excellent physical well-being.“
- Mismatching: „Ok, because it offers satisfying emotional well-being with excellent physical well-being, I imagine going for a walk is alright, you see?“

Da die Studie einem Within-Group-Design folgte, wurden alle Teilnehmer nacheinander mit allen Adaptionsstrategien konfrontiert. Zur Vermeidung von Reihenfolgeeffekten wurde die Reihenfolge der Adaptionsstrategien ausbalanciert und alle Teilnehmer wurden zufällig, aber möglichst gleichverteilt einer der möglichen Reihenfolgen zugeordnet. Alle Aussagen in den folgenden Fragebögen der Studie sollten mittels einer Likert-Skala von 1 = „trifft nicht zu“ bis 5 = „trifft zu“ bewertet werden.

Nach der Vorstellung der ersten Systempersönlichkeit, sollten die Teilnehmer jeden der drei Empfehlungstexte dahingehend bewerten, ob der Text (a) überzeugend und (b) vertrauenswürdig eingeschätzt wurde. Nach dem die Teilnehmer alle Beispiele gehört und bewertet hatten, erhielten sie einen weiteren Fragebogen mit Aussagen zur generellen Einschätzung des Prototypen. Dieser Fragebogen enthielt Aussagen wie „Ich würde den Roboter um Rat fragen.“, „Der Agent ist mir ähnlich.“ oder „Ich würde den Roboter gerne in Zukunft verwenden.“ Anschließend begann die Prozedur mit der nächsten Adaptionsstrategie bzw. Systempersönlichkeit von Neuem. Während der ganzen Studie gab es die Möglichkeit, Reetis Redegeschwindigkeit anzupassen, falls es Verständnisprobleme gab. Außerdem war es ebenfalls möglich Aussagen wiederholt abzuspielen.

Zum Abschluss der Studie wurden vergleichende Fragen hinsichtlich der Persönlichkeitsvarianten mit der höchsten Vertrauenswürdigkeit und der höchsten Überzeugungskraft gestellt. Zur Erleichterung der Einschätzung konnten sich die Studienteilnehmer die einzelnen Beispiele nochmals anhören. Außerdem wurde um Kommentare zum allgemeinen Eindruck des Systems gebeten.

Ergebnisse An der Studie nahmen 18 Personen im Alter von 25 bis 40 Jahren teil. Etwa 40% der Teilnehmer studierten an der Informatikfakultät der Universität Augsburg. Die restlichen Personen war in verschiedensten Bereichen berufstätig. Alle Teilnehmer waren Deutsch-Muttersprachler. Sie gaben an gute bis sehr gute Englischkenntnisse zu haben. Für die Analyse der Studienergebnisse konnten allerdings letztendlich nur 13 Personen berücksichtigt werden, da sich während der Studie herausstellte, dass fünf der Teilnehmer doch Schwierigkeiten hatten die englischen Äußerungen zu verstehen.

Die deskriptive Analyse der mittleren Bewertung für die Überzeugungskraft und Vertrauenswürdigkeit der Empfehlungstexte zeigte, dass das Mirroring für beide Faktoren bessere Bewertungen erhielt als die anderen beiden Strategien, die in etwa gleich gut abschnitten, siehe Abbildung 5.23.

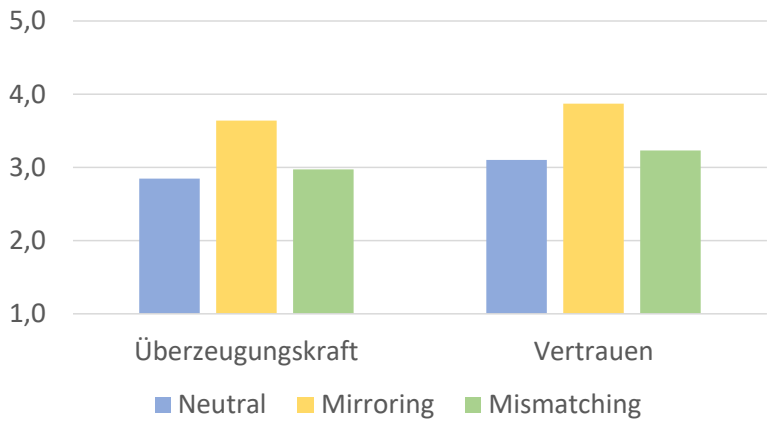

Abbildung 5.23: Mittlere Bewertung für die Überzeugungskraft und Vertrauenswürdigkeit der untersuchten Adaptionsstrategien

Ein ANOVA-Test mit Messwiederholung bestätigte, dass zwischen den Adaptionsstrategien sowohl für die Überzeugungskraft ($F(2,24)=4{,}480$; $p<0{,}5$) als auch für die Vertrauenswürdigkeit ($F(2,24)=4{,}513$; $p<0{,}5$) signifikante Unterschiede bestehen. Ein Bonferroni-Posthoc-Test zum paarweisen Vergleich ergab jeweils signifikant bessere Ergebnisse durch Mirroring als durch die Verwendung einer neutralen Persönlichkeit, siehe Tabelle 5.13. Zwischen Mirroring und Mismatching gab es keine signifikanten Unterschiede, aber starke Tendenzen zum Vorteil des Mirroring.

Tabelle 5.13: Wahrnehmung der Adaptionsstrategien (Abkürzungen: M = Mittelwert; SA = Standardabweichung; Sig = signifikant besser als)

	Überzeugungskraft			Vertrauenswürdigkeit		
	M	SA	Sig	M	SA	Sig
Neutrale Persönlichkeit	2,85	0,85		3,10	0,79	
Mirroring	3,64	0,63	1**	3,87	0,57	1*
Mismatching	2,97	0,94		3,23	0,95	
*signifikant mit p <0,05; **signifikant mit p <0,01; ***signifikant mit p <0,001						

Die durchschnittlichen Bewertungen für die weiteren Einschätzungen der drei Persönlichkeiten des Roboters sind in Abbildung 5.24 dargestellt. Es ergab sich nur für die Aussage „Ich würde den Roboter um Rat fragen." ein signifikanter Unterschied ($F(2, 24)=3{,}369$; $p<0{,}5$). Laut des paarweisen Vergleichs (Bonferroni-Posthoc-Test) würden die Studienteilnehmer den Roboter mit einer ähnlichen Persönlichkeit signifikant häufiger um Rat fragen, als den Roboter mit neutraler Persönlichkeit ($M(\text{Mirroring})= 3{,}38$; $SA(\text{Mirroring})=1{,}04$; $M(\text{neutral})= 2{,}77$; $SA(\text{neutral})=1{,}17$; $p<0{,}05$). Für die beiden anderen Aussagen ergaben sich keine signifikanten Unterschiede ($F_{aehnlich}(2, 15{,}24)=1{,}00$; $p=0{,}38$; $F_{Zukunft}(2, 24)=0{,}38$; $p=0{,}69$). Auffällig war allerdings, dass alle drei Adaptionsstrategien bei den Nutzern kein Gefühl von Ähnlichkeit hervorrufen konnten.

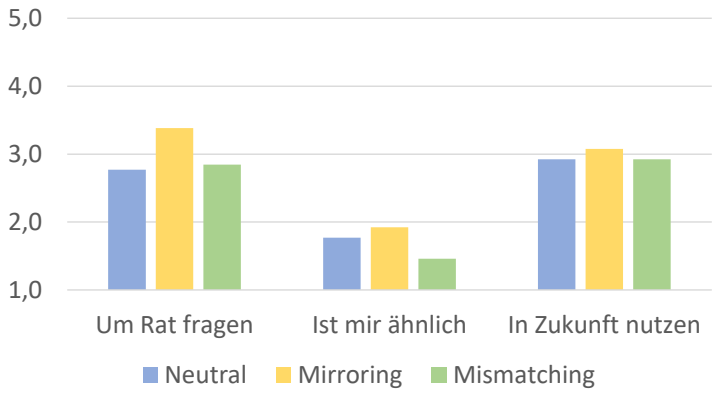

Abbildung 5.24: Mittlere Bewertung für die Aussagen „Ich würde den Roboter um Rat fragen.", „Der Agent ist mir ähnlich." und „Ich würde den Roboter gerne in Zukunft verwenden."

Die abschließenden vergleichende Fragen nach der Persönlichkeitsvariante mit der höchsten Vertrauenswürdigkeit und der höchsten Überzeugungskraft zeigten, dass Mirroring das größte Vertrauen hervorrief, während Mismatching als am überzeugendsten wahrgenommen wurde. Allerdings bestanden keine signifikanten Unterschiede ($F_{Vertrauen}(2, 24)=1{,}09$; $p=0.35$; $F_{Ueberzeugungskraft}(2, 24)=1{,}22$; $p=0.31$).

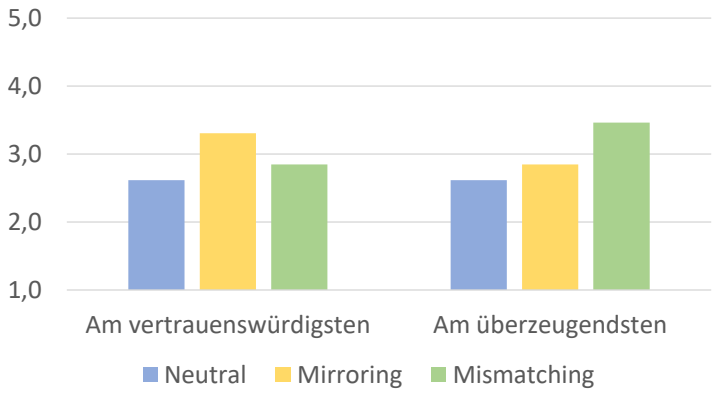

Abbildung 5.25: Mittlere Bewertung für die Frage „Welche Variante des Roboters fanden sie am vertrauenswürdigsten.“ und „Welche Variante des Roboters fanden sie am überzeugendsten.“

5.3.5 Diskussion

Der Vergleich der Adaptionsstrategien bestätigte, dass die Adaption der Empfehlungstexte sowohl die Vertrauenswürdigkeit (Hypothese 1) als auch die Überzeugungskraft (Hypothese 2) einer Empfehlung beeinflussen kann. Durch Mirroring konnten im Vergleich zur Verwendung einer neutralen Persönlichkeit beide Faktoren signifikant gesteigert werden. Allerdings entstand dieser Effekt lediglich bei der spontanen Bewertung der einzelnen Beispiel-Empfehlungen. Beim direkten Vergleich der Varianten am Ende der Studie erreichte Mirroring zwar wiederum die besten Bewertungen für die Vertrauenswürdigkeit. Am überzeugendsten wurde allerdings Mismatching eingeschätzt. Für beide Vergleiche waren die Unterschiede jedoch nicht statistisch signifikant. Es kann demzufolge festgehalten werden, dass viel darauf hindeutet, dass Mirroring tatsächlich zu einer erhöhten Vertrauenswürdigkeit führt (Hypothese 3). Für Hypothese 4, dass die Anwendung von Mismatching eine verbesserte Überzeugungskraft mit sich bringt, konnten jedoch keine eindeutigen Ergebnisse erzielt werden.

Ein Problem, das womöglich auch zu den geringeren Unterschieden hinsichtlich der Wahrnehmung von Mirroring und Mismatching beitrugt, ist, dass selbst manche der Nutzer, die sehr gut Englisch sprachen, die generierten Texte als „gewöhnungsbedürftig“ und „komisch“ bzw. die Aussprache des Roboters als „unnatürlich“ bezeichneten. Außerdem unterscheiden sich die Varianten der Empfehlungstexte nur in teilweise schwer einzuschätzenden Feinheiten. Womöglich hätte eine Studie in Deutsch mit klarer unterscheidbaren Empfehlungstexten (Satzlänge, Polarität der

Sätze) zu noch besseren Ergebnissen geführt. Auch ein stärkerer Fokus auf einzelne Persönlichkeitsmerkmale wie die Extraversion könnte die Effekte auf die Wahrnehmung der Empfehlungstexte weiter fördern.

Zu guter Letzt zeigten die Erfahrungen während der Studie, dass es schwierig ist, die Persönlichkeit eines Systems nur anhand dreier einzelner Aussagen zu beurteilen. Vor allem bei Systemen mit einer natürlichen Sprachausgabe sollten Empfehlungen in weiterführenden Arbeiten in Dialoge integriert werden. Dies würde den sozialen Faktor der Interaktion fördern und die Persönlichkeit der Aussagen stärker zum Ausdruck bringen. Außerdem könnte durch die Dialoge über die Zeit ein besseres Persönlichkeitsprofil der Nutzer erstellt werden.

5.4 Zusammenfassung

In diesem Kapitel wurde die Generierung von Empfehlungstexten in dreierlei Hinsicht untersucht und wichtige Erkenntnissen gewonnen.

Kulturbasierte Auswahl von Argumenten Zunächst wurde der Frage nachgegangen, ob der kulturelle Hintergrund von Menschen einen Einfluss auf die Überzeugungskraft von Argumenten für Energiesparempfehlungen hat. Darauf aufbauend wurde untersucht, ob Hofstedes Kulturmodell dafür geeignet ist, eine kulturbasierte Argumentauswahl zu realisieren. Die Ergebnisse einer Online-Studie zeigten, dass die Auswahl von Argumenten basierend auf den kulturellen Hintergründe der Nutzer sich positiv auf die Überzeugungskraft von Argumenten auswirken kann. Allerdings galten Themen wie Geld, Energiesparen und Umweltschutz interkulturell als wichtig und überzeugend. Des Weiteren wurde deutlich, dass eine reine Fokussierung auf Hofstedes Kulturdimensionen nicht ausreichend ist, um eine gute Personalisierung von Argumenten erreichen zu können.

Weitere Faktoren wie die Energiekultur der Nutzer könnten die Argumentauswahl weiter verbessern. Außerdem sollten persönliche Präferenzen der Nutzer berücksichtigt werden. Eine Argumentauswahl basierend auf individuellen Nutzerdaten und situativen Informationen hätte außerdem den Vorteil, dass die Argumente durch diese aktuellen Daten vielfältiger gestaltet werden könnten. Konkrete eingesparte Kosten, die individuelle Energiebilanz oder der aktuelle Spritpreis und das aktuelle Wetter könnten die Argumente ähnlich interessant machen, wie wechselnde Argumente, die zum Beispiel auch auf neue technische Errungenschaften oder wissenschaftliche Erkenntnisse hinweisen könnten.

Höflichkeitsstrategien in Empfehlungstexten Ziel der Untersuchungen war, herauszufinden, ob unterschiedliche Höflichkeitsstrategien zu unterschiedlichen Wahrnehmungen der Höflichkeit und der Überzeugungskraft von Empfehlungen führen würden. Die Ergebnisse zweier Studien zeigten, dass Höflichkeitsstrategien von den Nutzern tatsächlich unterschiedlich wahrgenommen werden. Dabei hingen die Höflichkeit und die Überzeugungskraft nicht voneinander ab. Formulierungen als Fragen oder als Ziele des Systems wurden zum Beispiel als sehr höflich, aber weniger überzeugend wahrgenommen. Dagegen galten direkte Kommandos zwar als am wenigsten höflich, dafür waren sie aber recht überzeugend. Höflich und überzeugend waren Formulierungen als gemeinsames Ziel und Anfragen bzw. Bitten.

Zusammengefasst können Höflichkeitsstrategien in beratenden Empfehlungssystemen also gezielt eingesetzt werden, um situative Ziele wie die Stärkung der Bindung zwischen Nutzer und System oder eine starke Überzeugungskraft des Empfehlungstextes zu erreichen. Die Ergebnisse der Studien sollten jedoch mit einem kompletten System nochmals untersucht werden, um Störvariablen wie nichtpersonalisierte Empfehlungen oder den Effekt einer neuen Technologie reduzieren und Langzeiteffekte untersuchen zu können.

Persönlichkeitsausprägungen von Formulierungen In diesem Kapitel wurde beschrieben, dass Nutzer einem System aufgrund seiner textuellen oder sprachlichen Ausgaben eine Persönlichkeit zuordnen. Dies geschieht unabhängig davon, ob diese Persönlichkeit von den Entwicklern des Systems intendiert ist oder nicht. Außerdem hat die jeweils wahrgenommene Persönlichkeit einen Einfluss darauf, wie sich Personen gegenüber einem System verhalten.

Um diesen Effekt für assistierende Empfehlungssysteme nutzbar zu machen, wurde ein Prototyp entwickelt, der das Framework PERSONAGE [Mairesse und Walker, 2007] nutzt, um Empfehlungstexte mit einer gewünschten Persönlichkeitsausprägung generieren zu können.

In einer Studie, in der der erstellte Prototyp eingesetzt wurde, konnte gezeigt werden, dass unterschiedliche Strategien zur Anpassung der Systempersönlichkeit an die jeweilige Zielperson zu unterschiedlichen Wahrnehmungen hinsichtlich der Vertrauenswürdigkeit und der Überzeugungskraft der Empfehlungstexte führen. Es wurden die Nutzung einer neutralen Persönlichkeit, Mirroring (Nutzung einer ähnlichen Persönlichkeit wie die der Zielperson) und Mismatching (Nutzung einer widersprüchlichen Persönlichkeit) vergleichen. Als vielversprechend stellte sich vor allem die Strategie des Mirroring heraus, die die Vertrauenswürdigkeit von Empfehlungstexten signifikant verbessern konnte. Im Bezug auf die Überzeugungskraft konnten keine eindeutigen Erkenntnisse gewonnen werden. Bei der Bewertung einzelner Empfehlungen schnitt Mirroring besser ab. Beim Gesamteindruck erzielte Mismatching etwas bessere Ergebnisse.

Die gemessenen Effekte in der Studie fielen jedoch durch komplizierte Sprachausgaben und große Ähnlichkeiten zwischen den Varianten der Empfehlungstexte geringer aus als erwartet. Extremere Persönlichkeitsausprägungen und deutlichere sprachliche Unterschiede wie die Satzlänge, die Polarität der Sätze oder Nachfragen beim Nutzer könnten einfacher zu erkennen sein und zu größeren Unterschieden bei der Einschätzung der Vertrauenswürdigkeit und Überzeugungskraft der Empfehlungstexte führen.

Weitere Einflussfaktoren in Empfehlungstexten In dieser Arbeit wurden sowohl der Inhalt (Argumente), als auch die Formulierung (Sprechart) von Empfehlungstexten untersucht. Die Struktur der Texte sowie die Organisation des Inhalts können allerdings ebenfalls einen Einfluss auf die Wahrnehmung der Empfehlung haben [Marcu, 1996, Reed und Long, 1997]. Deshalb sollten diese Faktoren in zukünftigen Arbeiten ebenfalls erforscht werden.

6 Proaktives Ausführen von Empfehlungen

Auch wenn Empfehlungen in vielen Situationen gerechtfertigterweise ausgewählt werden würden, besteht die Gefahr, dass sie, wenn sie zu häufig oder in unpassenden Situationen ausgesprochen werden, als störend und lästig wahrgenommen werden [Bader et al., 2010, Melguizo et al., 2007]. Das gilt in beratenden Empfehlungssystemen vor allem dann, wenn sich die empfohlenen Handlungen und Maßnahmen nur auf triviale Dinge wie zum Beispiel das Ausschalten einer Lampe im SavER-Szenario beziehen. In solchen Fällen könnten die Nutzer dann sogar das Gefühl haben, dass sie nicht beeinflussen können, wann und wie oft sie durch das System angesprochen werden. Die Frage, ob und wie ein System seine Nutzer bei der Durchführung ihrer eigentlichen Aufgaben unterbrechen sollte, um bestimmte Informationen zu präsentieren, wurde u.a. bereits von Horvitz und Kollegen erforscht [Horvitz, 1999, Horvitz et al., 2003]. Allerdings zielte ihre Forschung auf weniger kritische Szenarien wie Desktopanwendungen und E-Mailprogramme ab.

Eine Alternative zur Präsentation einer Empfehlung wäre die direkte Ausführung der Empfehlung durch das System. Auch Azaria und Hong [Azaria und Hong, 2016] erwähnen diese Option als wichtigen Bestandteil zukünftiger Empfehlungssysteme. Speziell im SavER-System würde es Sinn machen, einfache Maßnahmen wie das Ausschalten ungenutzter Geräte durch das System ausführen zu lassen, falls das Aussprechen einer Empfehlung unpassend ist. Diese Vorgehensweise hätte den Vorteil, dass eine Energieverschwendung verhindert würde, ohne dass die Nutzer gestört werden. Zusätzlich könnten die Nutzer durch diese kleinen Hilfen weiter von den Vorteilen eines energiesparenden Verhaltens überzeugt werden, wenn sie die positiven Folgen der Energiesparmaßnahmen erkennen.

Es stellt sich jedoch die Frage, in welchen Situationen autonomes Systemverhalten angemessen ist und in welchen nicht. Wie kann ein System einschätzen, ob es autonom handeln darf oder nicht? Ein Entscheidungskriterium könnten die affektiven und sozialen Reaktionen der Nutzer auf autonome Systemreaktionen sein. Diese Reaktionen wirken sich laut Reeves und Nass [Reeves und Nass, 1998] stark auf die Nutzerakzeptanz und die Interaktion mit einem System aus. Mitchell und Kollegen [Mitchell et al., 1994] berücksichtigten zum Beispiel zusätzlich zum situativen Vertrauen des Systems in seine eigene Entscheidungsfähigkeit auch das Vertrauen der Nutzer in die Entscheidungen des Systems, um zu entscheiden, ob ein System autonom Termine für seine Nutzer festlegen darf. Dass Nutzervertrauen mit dem Zutrauen in die Fähigkeiten eines Systems korreliert, ergab auch die Arbeit von Yu und Kollegen [Yu et al., 2017]. Sie zeigten auch, dass fehlerhafte Handlungen stärkere Änderungen des Vertrauens mit sich bringen als erfolgreiche Handlungen. Allerdings stabilisiert sich das Nutzervertrauen, wenn die Nutzer ein System besser kennen, so dass Fehler auch verziehen werden können.

Die Entscheidung pro oder kontra autonomes Systemverhalten hängt jedoch nicht nur vom Vertrauen in die Zuverlässigkeit des Systems ab. Aktionen autonomer und adaptiver Systeme stehen häufig im Widerspruch zu Kriterien wie Kontrollierbarkeit,

Transparenz oder Vorhersehbarkeit, die einen direkten Einfluss auf das Nutzervertrauen haben [Höök, 1997, Jameson, 2003], siehe Kapitel 2.3. Das folgende Beispiel soll die genannten Widersprüche anhand des SavER-Systems verdeutlichen.

Nehmen wir an, dass in einem Büro eine Lampe brennt, obwohl das Tageslicht für die Beleuchtung des Zimmers ausreicht. Wie soll das SavER-System in dieser Situation reagieren? Soll es davon ausgehen, dass die Mitarbeiter sich ihres Energieverbrauchs bewusst sind und selbst notwendige Maßnahmen ergreifen werden? Soll es das Licht selbst ausschalten? Oder soll es einen der Mitarbeiter durch das Aussprechen einer Empfehlungen auf dem Computerbildschirm oder dem Mobiltelefon auf den Missstand hinweisen und anbieten, das Licht auszuschalten? Im ersten Fall würde das System die Verantwortung für die nötige Maßnahme den Nutzern überlassen. Dadurch besteht die Gefahr, dass der Nutzen des Systems hinterfragt wird. Der zweite Ansatz würde zwar das Problem lösen und wäre bequem für die Nutzer. Er trägt allerdings auch die Gefahr, dass die Nutzer das Verhalten des Systems nicht verstehen und es womöglich sogar als zufällig handelnd wahrnehmen. Außerdem könnte es sein, dass sich die Nutzer bewusst für ein angeschaltetes Licht entschieden haben und sich bevormundet oder ihrer Kontrolle über das System beraubt fühlen. Im letzten Fall würde das System transparent handelnd und durch seine Nachfrage, den Nutzern die Entscheidung überlassen und somit ihre Autonomie respektieren. Allerdings könnten die Nutzer vom System genervt sein, wenn sie regelmäßige durch Empfehlungen bei ihrer Arbeit gestört werden.

In jedem Fall gilt: Trifft das System eine falsche Entscheidung, schadet dies dem Nutzervertrauen und damit auch der Nutzerakzeptanz. Im Wiederholungsfall könnten sogar eine weitere Nutzung des Systems abgelehnt werden.

Forschungsfrage Um Empfehlungssysteme wie SavER situative Entscheidungen über die Angemessenheit von Systemaktionen zu ermöglichen, wird in dieser Dissertation ein vertrauensbasierter Ansatz untersucht. Im Detail soll geklärt werden, wie das Nutzervertrauen gegenüber einem System modelliert werden kann, damit es bei der Entscheidungsfindung des Systems berücksichtigt werden kann. Außerdem soll erforscht werden, welche Faktoren das Nutzervertrauen und die Präferenzen der Nutzer für Systemaktionen im SavER-Szenario beeinflussen.

Kapitel 6.1 fasst weitere verwandte Arbeiten zusammen, um zu zeigen, in welcher Form die bereits vorgestellten Vertrauensdimensionen in Empfehlungssystemen und adaptiven Systemen berücksichtigt werden. Anschließend wird in Kapitel 6.2 das innerhalb des OC-Trust Projekts entwickelte User Trust Model (UTM) vorgestellt, das das Vertrauen der Nutzer in ein System und seine Fähigkeit zur Auswahl geeigneter Systemaktionen modellieren und darauf aufbauend situativ angemessene Entscheidungen treffen kann. Die Integration des UTM in ein SavER-System und eine anschließende Evaluation des Prototypen sind in Kapitel 6.3 zusammengefasst und können auch in zwei Veröffentlichungen nachgelesen werden [Hammer et al., 2014, Hammer et al., 2015c].

6.1 Vertrauensdimensionen in verwandten Arbeiten

In dieser Arbeit wurde ein Ansatz untersucht, der darauf beruht, die Wahrnehmung eines Empfehlungssystems und seiner Entscheidungen für die Nutzer so vertrauenswürdig wie möglich zu gestalten. Wie in Kapitel 2.3 beschrieben, haben verschiedene Faktoren, die *Vertrauensdimensionen*, einen Einfluss auf das affektive Nutzervertrauen: Nutzungskomfort, Transparenz, Kontrollierbarkeit, Privatsphäre, Zuverlässigkeit, Sicherheit, Glaubwürdigkeit sowie die Seriosität eines Systems und seiner Aktionen. All diese Dimensionen werden aus Sicht der Nutzer betrachtet. Um klar darzustellen, wie die Dimensionen in dieser Arbeit zu interpretieren sind, werden sie im Folgenden kurz beschrieben. Allerdings liegt der Fokus nur auf den Vertrauensdimensionen, auf die die aktuelle Situation und die Entscheidungen des Systems Auswirkungen haben: Nutzungskomfort, Transparenz, Kontrollierbarkeit, Privatsphäre. Die restlichen Faktoren beschreiben relativ konstante Systemeigenschaften und beeinflussen die Entscheidungsfindung des Systems nicht.

Transparenz Für Nutzer sollte immer sofort ersichtlich sein, warum ein System eine spezifische Anpassung vorgenommen hat [Gregor und Benbasat, 1999, Höök, 1997, Höök, 2000, Kay, 2006]. Auch im Falle von Empfehlungssystemen sollte möglichst einfach klar werden, warum die präsentierte Empfehlung in der aktuellen Situation und für die Nutzer persönlich relevant ist [Bader et al., 2011b, Cramer et al., 2008, Herlocker et al., 2000, Sinha und Swearingen, 2002]. Hier können Erklärungen sehr hilfreich sein [Gedikli et al., 2014, Pu und Chen, 2006, Tintarev und Masthoff, 2011, Zanker, 2012], siehe Kapitel 5.1. Allerdings ist es meist nicht notwendig, die technischen Hintergründe eines System im Detail zu erklären. Sind die Nutzer zum Beispiel Laien, können zu viele Details sogar abschrecken [Herlocker et al., 2000, Höök, 1997]. Einfache Erklärungen [Herlocker et al., 2000] oder graphische Darstellungen [Maes, 1994] können bereits ausreichen, um ein hohes Maß an Vertrauen und Nutzerakzeptanz zu erreichen.

Nutzungskomfort Ein wichtiger Bestandteil des Nutzungskomforts ist die Einfachheit der Nutzung, siehe Kapitel 2.2. Der Nutzungskomfort eines proaktiv handelnden Empfehlungssystems wird jedoch auch dadurch beeinflusst, ob es als aufdringlich oder ablenkend wahrgenommen wird [Bader et al., 2011b, Melguizo et al., 2007]. Ein Empfehlungssystem sollte daher nur mit seinen Nutzern interagieren, wenn es die Situation zulässt oder keine andere Wahl besteht. Gegebenenfalls sollten Aktionen auch autonom vom System durchgeführt werden können.

Nutzerkontrolle Die Nutzer eines Systems sollten immer die Kontrolle über die Systemaktionen innehaben. Ein Gefühl von Kontrollverlust schadet sowohl dem Vertrauen in das System als auch der Akzeptanz des Systems. Dies gilt vor allem dann, wenn die Bedürfnisse und Absichten der Nutzer im Gegensatz zu den Aktionen des Systems stehen [Jameson, 2003].

Schutz der Privatsphäre Das System sollte nur unbedingt benötigte private Daten sammeln und diese auf keinen Fall an andere preisgeben. Empfehlungen oder Anpassungen, die Rückschlüsse über Präferenzen und Daten der Nutzer zulassen, sollten demzufolge nie der Öffentlichkeit zugänglich gemacht werden [Wissner et al., 2014].

6.2 User Trust Model

Um das Nutzervertrauen mitsamt seiner Dimensionen in einen vertrauensbasierten entscheidungstheoretischen Ansatz für proaktiv handelnde Umgebungen integrieren zu können, wurde innerhalb des DFG-Projektes OC-Trust das *User Trust Model (UTM)* entwickelt. Es diente dazu drei wichtige Aufgaben des Vertrauensmanagements [Yan und Holtmanns, 2008] erfüllen zu können:

1. Einschätzung des aktuellen Vertrauens der Nutzer in das System
2. Fortwährende Überwachung der Entwicklung des Vertrauens
3. Automatische Ausführung geeigneter Systemaktionen, um das Vertrauen der Nutzer in kritischen Situationen mindestens zu erhalten, möglicherweise aber auch zu steigern

Für eine ausführlichere Diskussion verwandter Arbeiten und eine umfangreichere Dokumentation des UTM sei auf die Veröffentlichung *A User Trust Model for Automatic Decision-Making in Ubiquitous and Self-Adaptive Environments* [Hammer et al., 2016b] verwiesen.

Das User Trust Model wurde in Form eines *Bayes'schen Netzes* umgesetzt. Diese Netze sind gerichtete, azyklische Graphen bestehend aus Knoten, die Zufallsvariablen darstellen, und Kanten, durch die Einflüsse zwischen den Knoten durch abhängige Wahrscheinlichkeiten beschrieben werden können [Russell und Norvig, 2003]. Die Entscheidung, Bayes'sche Netze einzusetzen, beruhte auf der Überlegung, dass diese sich sehr gut eignen, um die folgenden typischen Eigenschaften des Nutzervertrauens abzubilden:

Subjektivität Verschiedene Personen nehmen die selben Situationen meist unterschiedlich wahr. Während manche Personen es zum Beispiel für kritisch halten, wenn ein System eigenständige Entscheidungen treffen kann, begrüßen andere Personen diese Fähigkeit. Generell bauen vertrauensselige Menschen auch schneller Vertrauen gegenüber einem Computersystem auf. In einem Bayes'schen Netz kann diese Subjektivität durch entsprechende Knoten und Wahrscheinlichkeitsverteilungen für unterschiedliche Grade von Vertrauen abgebildet werden.

Im UTM, siehe Abbildung 6.1, wurde die individuelle Bereitschaft zur Bildung von Vertrauen in ein technisches System im Knoten *Nutzernaturell* abgebildet. Dieses Naturell wird durch die individuelle (*Kompetenz*) und das generelle Vertrauen in technische Systeme (*Zutrauen*) abgebildet. Genau genommen werden im UTM

auch nicht die tatsächlichen Einflüsse der Situationen und Systemaktionen auf das Nutzervertrauen modelliert, sondern die Verbindung zwischen dem Vertrauen der Benutzer und ihrer subjektiven Wahrnehmung der Situationen und Systemaktionen. Die Darstellung eines Symbols, das ein geschlossenes Vorhängeschloss zeigt, hätte zum Beispiel keinerlei Einfluss auf das Vertrauen einer Person, wenn diese nicht das Hintergrundwissen besitzt, dass dieses Symbols auf die Sicherheit ihrer privaten Daten hinweist [Dhamija et al., 2006]. Wird im Folgenden von spezifischen Eigenschaften des System und seiner Systemaktionen geschrieben, ist also immer die subjektive Wahrnehmung dieser Eigenschaften durch die Nutzer gemeint.

Non-Determinismus Personen können vertrauenskritische Situationen als eher harmlos oder möglicherweise sogar überhaupt nicht wahrnehmen. Deshalb sind die Auswirkungen bestimmter Ereignisse auf das Nutzervertrauen meist nicht eindeutig vorherzusehen und feste Regelsätze zur Vorhersage von Nutzerreaktionen unbrauchbar. In Bayes'schen Netzen kann diese Unsicherheit durch abhängige Wahrscheinlichkeiten zwischen Knoten modelliert werden.

Im UTM könnte u.a. modelliert werden, dass eine moderate Transparenz einer Systemreaktion wahrscheinlich auch ein moderates Maß an Vertrauen zur Folge hat.

Vielseitigkeit Wie die Vertrauensdimensionen zeigen, hängt das Nutzervertrauen von einer Kombination mehrerer Facetten ab. Auch diese Eigenschaft kann relativ einfach in Bayes'schen Netzen abgebildet werden.

Abbildung 6.1 zeigt, dass innerhalb des UTM jede Vertrauensdimension durch einen eigenen Knoten repräsentiert und indirekt mit dem Knoten *Nutzervertrauen* verbunden ist. Die abhängigen Wahrscheinlichkeiten für die Einflüsse der Dimensionen auf das Vertrauen können entweder intuitiv oder für exaktere Vorhersagen durch empirische Daten abgeschätzt werden, siehe Kapitel 6.3.1. Sollte es zu einem späteren Zeitpunkt neue Erkenntnisse über die Zusammensetzung der Vertrauensdimensionen oder deren Einfluss auf das Nutzervertrauen geben, können diese durch eine Überarbeitung der entsprechenden Teile des Netzes einfach integriert werden.

Dynamische Entwicklung Die letzte wichtige Eigenschaft des Vertrauens ist, dass es sich mit der Zeit und abhängig von Erlebnissen ändern kann. Durch die Bündelung der Einflüsse einzelner Vertrauensdimensionen in zusätzlichen Knoten können solche Beziehungen in Bayes'schen Netzen ebenfalls gut abgebildet werden. In Abbildung 6.1 sind diese Knoten blau eingefärbt.

Der Arbeit von Lumsden folgend [Lumsden, 2009] wurde im UTM zwischen zwei Arten von Vertrauen unterschieden. *Initiales Vertrauen* entsteht durch die ersten Eindrücke, die eine Person von einem System gewinnt. Es kann zum Beispiel durch vorhandene Zertifikate (*Sicherheit*), ein seriöses Design (*Seriösität*) und zusätzliche Informationen, wie zum Beispiel Details über die verantwortliche Firma (*Glaubwürdigkeit*) aufgebaut werden. Im weiteren Verlauf der Nutzung des Systems kann anschließend weiteres Vertrauen gebildet werden (*Interaktiosbasiertes Vertrauen*).

Dieses Vertrauen wird zum einen von der *Zuverlässigkeit* des Systems und seinem Umgang mit privaten Daten der Nutzer (*Privatsphäre*) beeinflusst, zum anderen aber auch durch die *Qualität der Interaktion* selbst. Diese hängt wiederum von der wahrgenommenen *Transparenz*, der *Kontrollierbarkeit* und dem wahrgenommenen *Nutzungskomfort* ab.

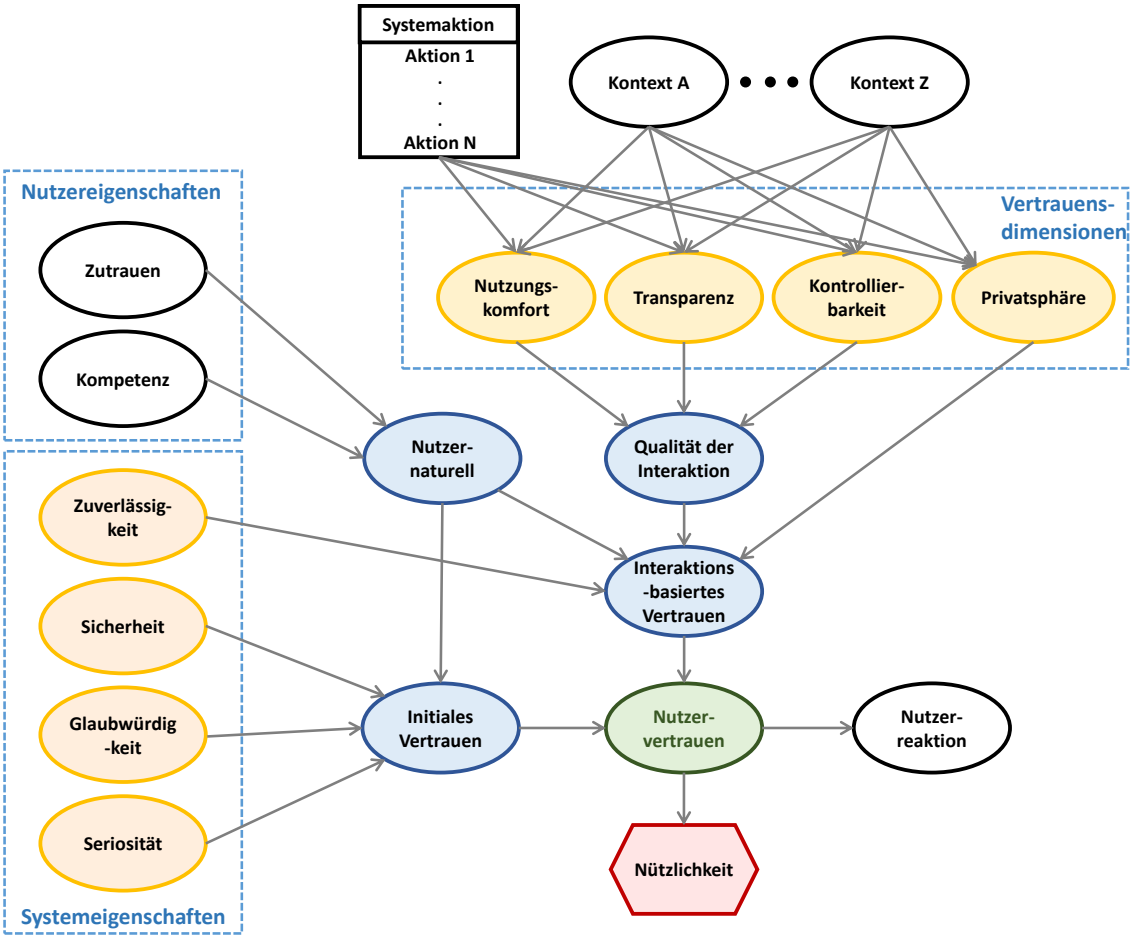

Abbildung 6.1: Generische Version des User Trust Models

Weitere Details der technischen Umsetzung Die Vertrauensdimensionen sind im UTM als versteckte Variablen konzipiert. Sie können weder direkt beobachtet noch gemessen werden, sondern hängen von der Wahrnehmung der sicht- und messbaren Kontextvariablen (*Kontext A* bis *Kontext Z*) und der jeweiligen *Systemaktion* ab. Beispiele für Kontextknoten wären der Status der Nutzer, die Anwesenheit anderer Personen (sozialer Kontext) oder Umwelteinflüsse wie die Helligkeit im Freien. Beispiele für Systemaktionen wären das Maskieren oder Verstecken persönlicher Daten oder die Generierung situativer Empfehlungen. Der Einfluss dieser systemabhängigen Knoten auf die Vertrauensdimensionen kann anschließend durch das UTM propagiert und letztendlich eine Abschätzung über ihren Einfluss auf das Nutzervertrauen getroffen werden.

Um basierend auf diesen Abschätzungen eine Auswahl aus den Systemaktionen treffen zu können, wurde das UTM zu einem sog. „Einflussdiagramm“ erweitert. Hierfür wurde der *Systemaktion*-Knoten zu einem Entscheidungsknoten umgewandelt, der mit einem zusätzlichen Knoten für die *Nützlichkeit* der Aktionen verbunden ist. Dieser Knoten berechnet die Nützlichkeit aller Systemaktionen in einer gegebenen Situation und gibt die Aktion mit der höchsten Nützlichkeit zurück. Da das Ziel des UTM ein möglichst großes Nutzervertrauen ist, wird die Nützlichkeit im UTM in direkter Abhängigkeit vom Nutzervertrauen bestimmt.

Da sein Kern generisch modelliert und mit empirischen Daten vorinitialisiert wurde [Bee et al., 2012, Leichtenstern et al., 2010], kann das UTM einfach für intelligente und proaktiv handelnde Systeme aller Art eingesetzt werden. Es müssen nur die systemabhängigen Kontexte und Systemaktionen definiert und ihr Einfluss auf die Dimensionen des interaktionsbasierten Vertrauens modelliert werden. Dies kann entweder durch intuitives Festlegen der abhängigen Wahrscheinlichkeiten oder durch die Sammlung empirischer Daten geschehen. Wie dies genau von statten gehen kann, zeigt die im Folgenden beschriebene Integration des UTM in ein SavER-System.

6.3 Evaluation im Anwendungsszenario SavER

Die Integration des UTMs in ein SavER-System wurde anhand eines intelligenten Büros untersucht. Cheverest und Kollegen befassten sich bereits im Rahmen einer „Smart Office“-Umgebung mit der Spannung zwischen proaktivem Systemverhalten und Nutzerkontrolle [Cheverst et al., 2005]. Sie untersuchten ebenfalls Techniken, um die Transparenz des Systems und die Nutzerkontrolle zu steigern. Obwohl sie zur Auswahl angemessener Systemaktionen nicht explizit das Vertrauen der Nutzer modellierten, ging ihr Ansatz bereits in diese Richtung. In dieser Arbeit sollte dieser Weg weitergegangen werden. Es wurde erforscht, ob es möglich ist, das Nutzervertrauen in ein SavER-System mit Hilfe des UTM zu modellieren und darauf aufbauend situativ und personalisiert Entscheidungen über die Angemessenheit von Empfehlungen und alternativer Systemaktionen zu treffen.

6.3.1 Integration des User Trust Models

Zwei Ursachen für verschwendete Energie in Büros sind unnötig eingeschaltete Lichter und Arbeitsrechner bzw. Bildschirme. Aus diesem Grund wurde je ein UTM für das Zimmerlicht und für die Bildschirme der Mitarbeiter erstellt. In Abbildung 6.2 ist ein Büro zu sehen, in dem der SavER-Prototyp installiert wurde. Es wird von zwei Mitarbeitern genutzt. Bei der Entwicklung und Evaluation des Prototypen wurde davon ausgegangen, dass ein Mitarbeiter die hauptsächliche Zielperson des Systems ist und der andere Mitarbeiter den sozialen Kontext darstellt.

Für die Erstellung und die Integration der UTMs in das System wurden die GeNIe Modellierungsumgebung und das Framework SMILE eingesetzt[23].

[23] http://genie.sis.pitt.edu/

Abbildung 6.2: Beispielhaftes Büro für den Einsatz des SavER-Prototypen

Entwurf des UTM-Licht Die Entscheidung, ob ein Zimmerlicht benötigt wird, hängt von der Helligkeit im Freien und der Anwesenheit einer oder mehrerer Person im Raum ab. Im UTM-Licht ist diese Kontextinformation durch die Kontextknoten *Helligkeit (außen)*, *Präsenz (Nutzer)* und *Präsenz (Kollege)* repräsentiert, siehe Abbildung 6.3. In Situationen, in denen das Zimmerlicht an- oder ausgeschaltet werden sollte, bieten sich dem System vier Möglichkeiten zu reagieren: (1) Automatisches An-/Ausschalten des Lichts, (2) Empfehlung zum An-/Ausschalten des Lichts über das Smartphone der Zielperson, (3) Empfehlung zum An-/Ausschalten des Lichts über den Bildschirm der Zielperson oder (4) Keine Reaktion

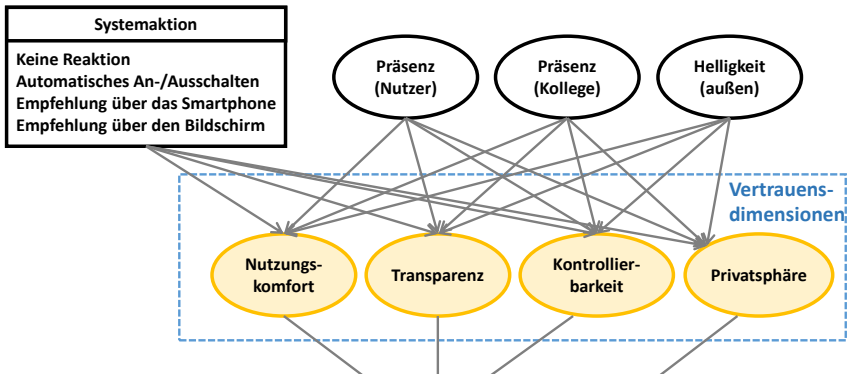

Abbildung 6.3: Ausschnitt aus dem UTM-Licht

Entwurf des UTM-Bildschirm Ob der Bildschirm einer Person benötigt wird, hängt einzig und allein vom aktuellen Status dieser Person ab, siehe Abbildung 6.4. Anders als im UTM-Licht reicht es allerdings nicht, nur die Anwesenheit zu überprüfen. Eine Person kann ihren Bildschirm aktiv nutzen. Sie kann auch direkt vor dem Bildschirm sitzen, ihn aber nicht nutzen, da sie in andere Aktivitäten wie zum Beispiel Lesen vertieft ist. Außerdem könnte sie sich zwar im Büro befinden, aber nicht am Schreibtisch sitzen. Ist ein An- bzw. Abschalten des Bildschirms nötig, bleiben dem System die gleichen Optionen wie bei der Adaption des Lichts. Allerdings entfällt logischerweise die Option, eine Empfehlung auf dem betroffenem Bildschirm anzuzeigen, wenn dieser nicht von der Zielperson beachtet wird.

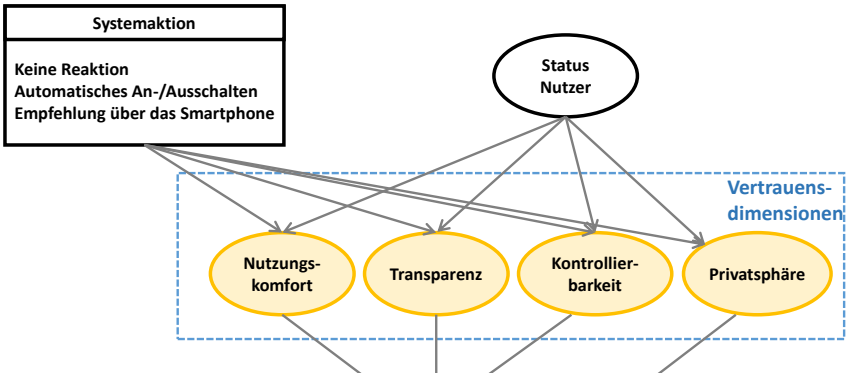

Abbildung 6.4: Ausschnitt aus dem UTM-Bildschirm

Eine Zusammenfassung der verwendeten Kontexte und der möglichen Systemaktionen für beide UTMs ist in Tabelle 6.1 zu finden.

Systemarchitektur Abbildung 6.5 zeigt die Systemarchitektur des entwickelten Prototypen. Zur Sammlung der benötigten Kontextdaten wurden verschiedene Arduino-Sensoren[24] installiert. Dazu gehörten Lichtsensoren, um die Helligkeit im Freien zu messen, Ultraschall-Sensoren, um die Anwesenheit der Personen an ihren Schreibtischen zu erkennen, und sog. „Flex-Sensoren“, mit denen erkannt werden konnte, ob die Tür des Büros geschlossen oder geöffnet war. Zur Verwendung der letzteren Sensoren ist anzumerken, dass in den Büros, in denen der Prototyp installiert wurde, eine „Open Door“-Politik herrscht. Das heißt, sobald sich einer oder mehrere Mitarbeiter im Büro befinden, bleibt die Tür solange geöffnet, bis alle Mitarbeiter das Büro verlassen haben. Somit kann, auch wenn sich keine Person an einem der Schreibtische befindet, davon ausgegangen werden, dass sich bei einer geöffneten Tür immer noch mindestens eine Person im Raum befindet. Die gesammelten Rohdaten wurden auf einem zentralen Server analysiert und die aggregierten Kontextinformationen in die beiden UTMs eingegeben. Je nachdem, ob das System automatisch ein Geräte kontrollieren oder eine Empfehlung aussprechen wollte, standen entweder ein HomeMatic-System[25] mit fernbedienbaren Steckdosen

[24] http://arduino.cc/

[25] http://www.homematic.com/

oder eine Verbindung zum Senden von Nachrichten an das Smartphone oder den Bildschirm der Zielperson zur Verfügung.

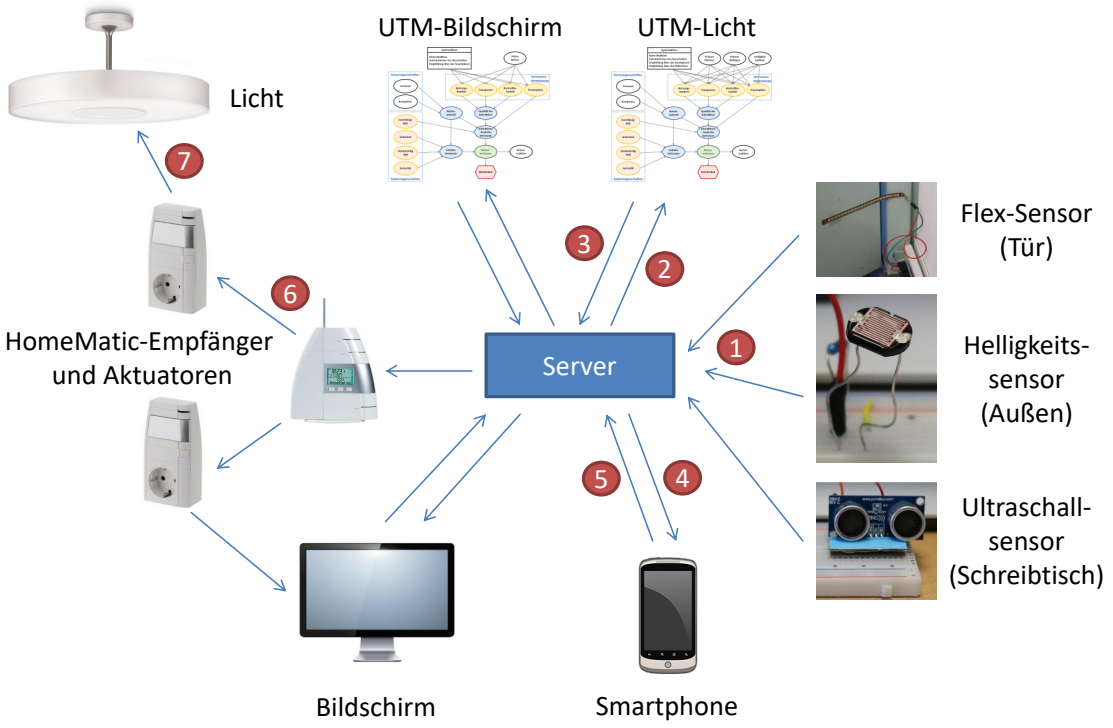

Abbildung 6.5: Architektur und Kontextwertschöpfungskette des SavER-Prototypen

Kontextwertschöpfungskette Das folgende Beispiel stellt einen typischen Ablauf einer Entscheidungsfindung innerhalb des entwickelten Prototypen vor. Die einzelnen Schritte sind in Abbildung 6.5 durch eingekreiste Zahlen dargestellt.

Das Szenario beginnt morgens. Draußen ist es noch dunkel. Der Nutzer ist alleine in seinem Büro und arbeitet am PC. Das Licht ist an. Das System ist sich dieser Situation bewusst, da es regelmäßig die Kontextdaten der Sensoren abruft (1). Ein Eingreifen ist aktuell nicht nötig. Etwas später wird es draußen heller. Da der Nutzer in seine Arbeit vertieft ist, fällt ihm das nicht auf. Das System registriert die Änderung und aktualisiert die Kontextinformation im UTM-Licht (2). Es muss entschieden werden, welche der vier Systemaktionen, siehe Tabelle 6.1, die am besten geeignete ist. Nehmen wir nun einmal an, dass das UTM für alle vier möglichen Systemreaktionen das resultierende Nutzervertrauen abgeschätzt und sich für das Senden einer Empfehlung auf das Smartphone des Nutzers entschieden hat. Diese Entscheidung wird nun dem zentralen Server mitgeteilt (3) und dieser sendet eine entsprechende Nachricht an das Smartphone. Der Nutzer hat nun die Möglichkeit, die empfohlene Systemaktion zu bestätigen oder abzulehnen (4). Bestätigt er die Empfehlung, wird diese Information wiederum über den Server (5) an das HomeMatic-System weitergeleitet (6), das anschließend das Licht ausschaltet (7).

Initialisierung Da nur der Kern des Bayes'schen Netzes vorinitialisiert war, wurden zunächst in zwei Onlineumfragen Daten gesammelt, mit denen die UTMs für das Licht und den Bildschirm fertig initialisiert werden konnten. In beiden Umfragen wurden den Teilnehmern textuell typische Situationen im Büroalltag beschrieben, in denen das System eine Entscheidung treffen müsste. Für jede dieser Situationen wurden die möglichen Systemreaktionen zur Steigerung der Energieeffizienz vorgestellt. In Tabelle 6.1 sind die Kontextvariablen mit den möglichen Zuständen sowie die Systemreaktionen für beide Geräte aufgelistet.

Tabelle 6.1: Berücksichtigte Kontextinformationen und mögliche Systemreaktionen

Gerät	**Situation**			**System-reaktion**
	Nutzer Status	**Sozialer Kontext**	**Helligkeit (Außen)**	
Bildschirm	1) Aktiv am PC 2) Inaktiv am PC 3) Abwesend vom PC 4) Abwesend	- - - - -	- - - - -	a) Automatisches An-/Abschalten b) Empfehlung per Smartphone c) Keine Reaktion
Licht	1) Im Raum 2) Abwesend	1) Kollege im Raum 2) Kollege abwesend	1) Dunkel 2) Hell	a) Automatisches An-/Abschalten b) Empfehlung per Smartphone c) Empfehlung per Bildschirm c) Keine Reaktion

Das Ziel der Umfrage war, herauszufinden, welchen Einfluss die möglichen Systemaktionen in den präsentierten Situationen auf das Nutzervertrauen haben. Aus diesem Grund sollten die Teilnehmer für jede Situation für jede Reaktion ihre Wahrnehmung der Transparenz und Kontrollierbarkeit sowie des Nutzungskomforts und des resultierenden Vertrauens in das System einschätzen. Die dazugehörigen Aussagen (Q1-Q4) sollten mittels einer 5er-Likert-Skala bewertet werden.

- Q1: Ich habe verstanden, warum das System auf diese Weise reagiert hat.
- Q2: Ich hatte Kontrolle über das System.
- Q3: Ich halte das System für bequem bedienbar.
- Q4: Ich halte das System für vertrauenswürdig.

Obwohl Verletzungen der Privatsphäre nie komplett ausgeschlossen werden können, waren sie in den untersuchten Situationen äußerst unwahrscheinlich. Aus diesem Grund wurde diese Vertrauensdimension, anders als in vorhergehenden Untersuchungen mit anderen Systemen [Wissner et al., 2014], nicht explizit abgefragt.

Insgesamt füllten sieben Frauen und neun Männer die Umfrage bezüglich des UTM-Licht aus. Für das UTM-Bildschirm konnten 22 Teilnehmer (10 Frauen, 12 Männer) gewonnen werden. Die Teilnehmer der Umfragen waren zwischen 24 und 51 Jahre alt (⌀ 28). Mit Hilfe der gewonnenen Daten konnten anschließend die abhängigen Wahrscheinlichkeit hergeleitet werden, die für jede Situation den Einfluss der Systemreaktionen auf die Vertrauensdimensionen beschrieben.

6.3.2 Evaluation

Bereits die Ergebnisse der Onlineumfrage können Auskunft darüber geben, wie Nutzer die Entscheidungen des SavER-Systems in bestimmten Situationen bewerten würden. Allerdings birgt diese Art der Datensammlung den Nachteil, dass die Bewertungen der Nutzer dadurch beeinflusst werden könnten, dass sie die Situation nicht selbst erlebt haben. Deshalb wurde mit einer Studie in einer realen Umgebung evaluiert, inwiefern das mit den Daten der Onlineumfrage initialisierte UTM tatsächlich dazu in der Lage ist, das Vertrauen und die Präferenzen der Nutzer vorherzusagen. Dabei lag weniger der technische Aspekt des Ansatzes im Fokus, als die persönlichen Erfahrungen mit und die Akzeptanz der Nutzer gegenüber der intelligenten SavER-Umgebung.

Hypothesen

1. Die Entscheidungen des UTM haben einen positiven Einfluss auf die Vertrauensdimensionen.

2. Die Entscheidungen des UTM beeinflussen das Vertrauen der Nutzer positiv.

3. Die vom UTM ausgewählten Systemaktionen stimmen mit den Präferenzen der Nutzer überein.

Experimenteller Aufbau Während der Studie wurde anhand einiger Aufgaben und Situationen ein typischer Alltag in einem Büro simuliert, das von zwei Mitarbeitern genutzt wird. Die Teilnehmer der Studie agierten als hauptsächliche Nutzer bzw. Zielpersonen des Systems. Der Evaluator übernahm die Rolle des Kollegen und beeinflusste dadurch den sozialen Kontext. Um sicherzustellen, dass alle Teilnehmer die Studie unter den selben Bedingungen durchführten, wurde das Büro verdunkelt und die Änderungen der Helligkeit im Freien durch eine Lampe simuliert. Außerdem wurde der Lichtsensor je nach gewünschter Situation ab- und aufgedeckt.

Durchführung der Evaluation Zu Beginn sollten die Teilnehmer einen Fragebogen ausfüllen, der demographische Daten, bisherige Erfahrungen mit Smart-Home-Systemen und das generelle Vertrauen in Computersysteme erfragte. Außerdem wurden sie gefragt, ob sie sich selbst als vertrauensselig beschreiben würden.

Anschließend erhielten alle Personen eine kurze Einführung in das Setting und das Szenario. Dann folgte mit dem Ausgangspunkt auf dem Flur die erste Aufgabe (Betreten des Büros), die zu einer Systemreaktion bzgl. des Lichts führte. Diese Reaktion sollten die Teilnehmer mittels eines kurzen Fragebogens bewerten. Dieser Fragebogen enthielt zum einen die Fragen Q1-Q4, die schon im Onlinefragebogen zum Einsatz kamen. Außerdem sollten die Teilnehmer angeben, welche Systemreaktion sie in der gegebenen Situation bevorzugen würden. Hierfür enthielt jeder Fragebogen pro möglicher Aktion eine Aussage, die mittels einer 5er-Likert-Skala von „auf keinen Fall“ bis „auf jeden Fall“ bewertet werden sollte. Diese Aussagen setzen sich aus einer kurzen Zusammenfassung der Situation und der jeweiligen Aktion zusammen. Für die erste Aufgabe lauteten die Aussagen dementsprechend: „Wenn ich mein Büro betrete, bevorzuge ich es, wenn...

- P1: ...das System nicht reagiert."
- P2: ...das System das Licht automatisch einschaltet."
- P3: ...das System mir über eine Nachricht auf mein Smartphone vorschlägt, das Licht automatisch einzuschalten."
- P4: ...das System mir über eine Nachricht auf meinem Bildschirm vorschlägt, das Licht automatisch einzuschalten."

Im selben Schema wie bei der ersten Aufgabe verlief auch die weitere Studie. Alle Aufgaben, die dazugehörigen Kontexte und die in der jeweiligen Situation ausgewählten Reaktionen des Systems sind in Tabelle 6.2 zusammengefasst. Damit sich die Teilnehmer besser in die Situationen hineinversetzen konnten, wurden alle Aktionen und Kontextänderungen in eine fortlaufende Erzählung integriert.

Zum Abschluss der Studie erhielten die Teilnehmer einen letzten Fragebogen. Darin sollten sie angeben welche Dinge ihnen an dem System gefallen bzw. nicht gefallen hatten. Außerdem sollten sie weitere Fragen beantworten, die auf dem TAM basierten, siehe Kapitel 2.2.

Ergebnisse Insgesamt nahmen sechs Frauen und 18 Männer im Alter von 23 bis 33 ($\oslash$ 26) an der Studie teil. 88% der Teilnehmer studierte oder arbeitete im Bereich Informatik. Der Rest verteilte sich über unterschiedliche Studiengänge und Berufe.

Da alle Aussagen in den Fragebögen mittels einer 5er-Likert-Skala bewertet werden sollten, wurde eine Bewertung von 3 als neutrale Haltung interpretiert. Höhere Bewertung wurden als Zustimmung und niedrigere Bewertung als Ablehnung der jeweiligen Aussage gewertet.

Nur fünf der 24 Teilnehmer gaben an, dass sie viel bis sehr viel Erfahrung mit Technologien zur Haussteuerung hatten. Beispiele für genutzte Technologien waren automatische Timer und Jalousiesteuerungen. Bis auf eine weitere Person, die ihre Erfahrungen neutral bewertete, hatten alle anderen Teilnehmer wenig bis keine Erfahrung mit Systemen zur Hausautomatisierung.

Tabelle 6.2: Aufgaben, Situationen und Systemreaktionen in der Nutzerstudie (Abkürzungen: Nichts = Keine Systemreaktion; Auto = Automatisches Ein/Ausschalten; ESP = Empfehlung auf Smartphone; EB = Empfehlung auf Bildschirm)

Aufgabe	**Situation**			**Systemreaktion**	
	Nutzer-zustand	**Sozialer Kontext**	**Helligkeit (Außen)**	**Licht**	**Bildschirm**
1. Betreten des Büros	Im Raum	Kollege abwesend	Dunkel	ESP	
2. An den Schreibtisch setzen	Aktiv am PC				Auto
Es wird draußen hell.					
3. Präsentation auf Fehler überprüfen.			Hell	EB	
Ein Kollege betritt den Raum und setzt sich an seinen Schreibtisch.					
4. Hole Buch X aus dem Schrank.	Abwesend vom PC	Kollege im Raum			Nichts
5. Komm zurück und lies Kapitel Y.	Inaktiv am PC				Auto
6. Ergänze Folie über Thema Y.	Aktiv am PC				Auto
Es wird draußen dunkel.					
			Dunkel	ESP	
Der Kollege verlässt den Raum.					
7. Verlasse das Büro und schließe die Tür	Abwesend	Kollege abwesend		ESP	Auto

Zur Einschätzung der Vertrauensseligkeit der Teilnehmer enthielt der erste Fragebogen zwei generelle Aussagen und eine Aussage mit Bezug zu Computersystemen:

1. Ich handle nach dem Vorsatz „Vertrauen ist gut, Kontrolle ist besser."
2. Ich bin übermäßig vertrauensselig.
3. Bei den meisten Computersystemen kann man sich sicher sein, dass sie tun was sie sollen.

Lediglich einer der Studienteilnehmer widersprach der ersten Aussage. Der größte Teil der Befragten (63%) stimmte der Aussage zu, während das restliche Drittel eine neutrale Bewertung abgaben. Für die beiden anderen Aussagen äußerte sich jeweils ein Drittel zustimmend, ablehnend und neutral.

Die Entscheidungen des Systems hinsichtlich des Lichts wurden von allen Teilnehmern konsistent mit hohen Bewertungen (4 oder 5) für die Vertrauensdimensionen

Transparenz, *Kontrollierbarkeit*, *Nutzungskomfort* und auch das *Vertrauen* selbst bewertet, siehe Tabelle 6.3. Allerdings bemängelten einige Nutzer, dass ihnen eine Rückmeldung des Systems fehlte, wenn sie das Büro verlassen hatten. Die Ungewissheit, ob das Licht wirklich ausgeschaltet wurde, beeinflusste auch das Vertrauen gegenüber dem System negativ.

Trotz der guten Bewertungen für die ausgewählten Systemreaktionen, präferierten die meisten der Nutzer in den meisten Situationen andere Systemreaktionen, siehe Tabelle 6.3-SRpräf. Bei diesen Situation handelte es sich vornehmlich um Situationen, in denen das System über das Smartphone Empfehlungen aussprach. Passend dazu äußerten einige Nutzer, dass sie die Nutzung des Smartphones in den Situationen umständlich fanden. Gründe dafür waren, dass es entweder außer Reichweite z.B. in der Tasche lag oder weil die Nutzer ihre Arbeit unterbrechen mussten, um das Smartphone nehmen und die Nachricht lesen zu können. Deshalb präferierten die Nutzer vor allem das automatische Ein- und Ausschalten des Lichts, das in ihren Augen das System weniger aufdringlich erscheinen ließ.

Die favorisierten Systemreaktionen für den Bildschirm stimmten dagegen bei den meisten Nutzern mit den vom UTM ausgewählten Systemreaktionen überein, siehe Tabelle 6.3-SRpräf. Betrachtet man die Entscheidung, nicht zu reagieren, ebenfalls als autonome Entscheidung über den aktuellen Zustand des Geräts, hätten ein Großteil der Studienteilnehmer die Entscheidungen bzgl. des Bildschirms in allen Situationen komplett dem System überlassen. Allerdings ist in Tabelle 6.3 zu sehen, dass dieses autonome Verhalten des Systems zu schlechteren Bewertung hinsichtlich der *Kontrollierbarkeit* und des *Nutzervertrauens* führte. Während das Nutzervertrauen immerhin noch einigermaßen positiv bewertet wurde (zwischen 3,5 und 4), waren die Durchschnittsbewertungen für die Kontrollierbarkeit nur noch mittelmäßig (zwischen 2,5 und 3,5). Als Gründe gaben die Nutzer wiederum fehlende Rückmeldungen nach Verlassen des Arbeitsplatzes an. Außerdem fehlte einigen Nutzern eine Authentifizierungsmechanismus beim Anschalten des Bildschirms oder eine Möglichkeit zur Aktivierung und Deaktivierung des autonomen Verhaltens des Bildschirms.

Die abschließenden Fragen unterstrichen die bisher vielversprechenden Resultate und deuteten auf eine Akzeptanz des Systems durch einen Großteil der Studienteilnehmer hin. Hinsichtlich der wahrgenommenen Nützlichkeit zeigten sich die meisten der Teilnehmer mit dem System zufrieden (83%; M=3,96; SA=0,68) und bestätigten, dass sie das System bei der Reduzierung des Energieverbrauchs unterstützte (96%; M=4,71; SA=0,54). Auch für die wahrgenommene Einfachheit der Nutzung konnten gute Ergebnisse erzielt werden. Das Verhalten des Systems wurde als angemessen (88%; M=4.38; SA=0.70) und transparent (100%; M=4.96; SA=0.20) eingestuft. Die etwas schlechteren, aber immer noch guten Bewertungen hinsichtlich der Aufdringlichkeit des Systems (58%; M=3.71; SA=1.10) lassen sich hauptsächlich auf die Nutzung des Smartphones zurückführen. Weitere Ergebnisse zeigten, dass sich die Nutzer weder gestört (75%; M=2.00; SA=1.00), eingeschränkt (88%; M=1.83; SA=1.07) noch überwacht (63%; M=2.33; SA=1.18) fühlten.

Tabelle 6.3: Nutzerbewertungen für die Vertrauensdimensionen und das Vertrauen in Abhängigkeit von der ausgeführten Systemreaktion (SR) und präferierte Systemreaktionen (SRpräf) (Abkürzungen: M = Mittelwert; SA = Standardabweichung; K = Kollege; Nichts = Keine Systemreaktion; Auto = Automatisches Ein/Ausschalten; ESP = Empfehlung auf Smartphone; EB = Empfehlung auf Bildschirm)

Situation	SR	Transparenz	Kontrollierbarkeit	Nutzungskomfort	Nutzer-Vertrauen	SRpräf (Anteil Nutzer)
Live-Studie - Gerät: Licht						
Anwesend, Dunkel, K abwesend	ESP	M=5,00 SA=0,00	M=4,46 SA=0,87	M=4,17 SA=0,99	M=4,25 SA=0,72	Auto (75%)
Hell, K abwesend	EB	M=4,92 SA=0,28	M=4,58 SA=1,04	M=4,63 SA=0,70	M=4,42 SA=0,64	EB (79%)
Dunkel, K im Raum	ESP	M=4,67 SA=0,75	M=4,25 SA=1,20	M=4,13 SA=0,97	M=4,13 SA=0,78	EB (58%)
Abwesend, Dunkel, K abwesend	ESP	M=4,92 SA=0,28	M=4,13 SA=1,27	M=4,29 SA=1,10	M=3,92 SA=0,86	Auto (67%)
Live-Studie - Gerät: Bildschirm						
Aktiv am PC	Auto	M=4,83 SA=0,47	M=2,83 SA=1,31	M=4,58 SA=0,57	M=3,75 SA=1,09	Auto (54%)
Abwesend vom PC	Nichts	M=3,79 SA=1,35	M=2,79 SA=1,55	M=4,13 SA=1,20	M=3,75 SA=1,13	Nichts (88%)
Inaktiv am PC	Auto	M=4,58 SA=0,91	M=2,50 SA=1,29	M=4,00 SA=1,08	M=3,63 SA=0,95	Auto (71%)
Abwesend	Auto	M=5,00 SA=0,00	M=3,46 SA=1,44	M=4,46 SA=1,15	M=3,88 SA=0,88	Auto (79%)

Zwischenfazit Trotz der vielversprechenden Ergebnisse der Studie gab es auch Dinge, die verbessert und genauer untersucht werden sollten. Das UTM-Bildschirm war gut darin, die Nutzerpräferenzen vorherzusehen. Das war beim UTM-Licht, obwohl die Vertrauensdimensionen und das Nutzervertrauen für die durchgeführten Systemaktionen sehr gut eingeschätzt wurden, nicht der Fall. Beim UTM-Bildschirm fielen dafür überraschenderweise die Bewertungen der Kontrollierbarkeit und des Vertrauens schlechter aus.

6.3.3 Erweiterte Evaluation

Da die Teilnehmer der Studie die alternativen Systemaktionen nicht hinsichtlich der Vertrauensdimensionen und dem Nutzervertrauen bewerten konnten, war eine genauere Analyse der Zusammenhänge zwischen den Nutzerpräferenzen und den Vertrauensdimensionen bzw. dem Nutzervertrauen nicht möglich. Um die Kriterien der Nutzer für die Wahl bestimmter Systemaktionen besser verstehen zu können, wurde deswegen im gleichen Setting wie bei der Studie eine Umfrage - zur besseren Unterscheidung zur Onlineumfrage ab sofort „Live-Umfrage“ genannt - durchgeführt. Während der Live-Umfrage konnten die Nutzer, im Gegensatz zur Onlineumfrage, alle Situationen und alle Systemreaktionen am eigenen Körper erleben. Das Konzept sowie die Resultate der Umfrage werden im Folgenden beschrieben.

Experimenteller Aufbau der Umfrage Das Ziel der Live-Umfrage war das Sammeln von Nutzerbewertungen für die Vertrauensdimensionen und das Vertrauen für alle Kombinationen aus Situationen und Systemreaktionen in einer realen Umgebung. Um vergleichbare Resultate zu erhalten, wurde der Aufbau der vorangegangenen Studie übernommen, siehe Kapitel 6.3.2. Allerdings war das UTM deaktiviert und es wurden alle Aktionen in allen Situationen präsentiert.

Zu Beginn der Befragung sollten die Teilnehmer wiederum demographische Fragen beantworten. Dann erhielten sie eine kurze Einführung in das Szenario sowie in das System mit allen Systemaktionen. Anschließend sollten die selben Aufgaben und Situationen wie in der Studie durchlaufen werden, siehe Tabelle 6.2. Für alle während der Studie präsentierten Systemreaktionen sollte die subjektive Wahrnehmung der Kriterien *Transparenz*, *Kontrollierbarkeit*, *Nutzungskomfort* und *Nutzervertrauen* bewertet werden. Dies geschah mit Hilfe der Aussagen Q1-Q4, die bereits in der Onlineumfrage und der Studie zum Einsatz kamen, siehe Kapitel 6.3.1.

Nachdem die Teilnehmer in einer Situation alle Systemaktionen erlebt und einzeln bewertet hatten, sollten sie außerdem angeben, welche der Aktionen sie bevorzugen würden. Allerdings taten sie dies nicht durch die Auswahl einer der Aktionen. Stattdessen sollten sie für jede Aktion die folgende Aussage auf einer 5er-Likert-Skala bewerten: „Ich würde die Systemaktion ... bevorzugen.“. Auf diese Weise konnten mehr Informationen zur Analyse der Präferenzen der Nutzer gewonnen werden.

Ergebnisse Für die Live-Umfrage konnten insgesamt 10 Personen (zwei Frauen, acht Männer) im Alter von 23 bis 35 (⌀ 28) akquiriert werden. Für den Vergleich der Entscheidungen des UTM in der Studie mit den Entscheidungen der Teilnehmer der Live-Umfrage wurden aus den gesammelten Bewertungen der Live-Umfrage für jede Situation die am meisten favorisierte und die als am meisten vertrauenswürdig wahrgenommene Systemaktion bestimmt.

Im Vergleich zur Studie steigerte sich in der Live-Umfrage die durchschnittliche Übereinstimmung der Nutzerpräferenzen mit der Reaktionsauswahl des UTM-Bildschirm von 73% auf 90%. 80% der Befragten brachten diesen Reaktionen auch das größte Vertrauen entgegen. Dies war bei einer durchschnittlichen Bewertung von 3,75 für das Vertrauen in die ausgewählten Systemaktionen in der Studie jedoch nicht erwartet worden. Allerdings wurden die Teilnehmer der Studie nicht nach ihrem Vertrauen in die alternativen Systemaktionen gefragt, weshalb dort kein Vergleich der Systemaktionen möglich war. Insgesamt zeigte sich in der Live-Umfrage hinsichtlich des UTM-Bildschirm für die Präferenzen und das Vertrauen der Nutzer eine übereinstimmende Tendenzen.

Wie bereits in der Studie, widersprachen sich auch in der Live-Umfrage die Vorlieben der Nutzer und die Entscheidungen des Systems bezüglich des Lichts in vielen Fällen. Während in der Studie noch ca. ein Drittel der Teilnehmer die gleiche Auswahl wie das UTM-Licht getroffen hätten, waren es in der Live-Umfrage nur noch 18%. Nichtsdestotrotz hatten auch beim Licht 80% der Nutzer das größte Vertrauen in die vom System ausgewählten Systemaktionen. Das UTM-Licht war also, wie in der Studie (durchschnittliches Vertrauen 4,18), dazu in der Lage vertrauenswürdige Entscheidungen zu treffen.

Die Lehre aus diesen Resultaten ist, dass das Vertrauen nicht das alleinige Entscheidungskriterium für Nutzer zu sein scheint, wenn sie ihre Präferenzen festlegen. Außerdem scheinen manche Vertrauensdimensionen die Präferenzen der Nutzer stärker zu beeinflussen als andere. So deuteten die Ergebnisse an, dass vielen Teilnehmern der Nutzungskomfort wichtiger ist als die Nutzerkontrolle. Um diese Hypothesen zu bestätigten, wurden die Daten der beiden Umfragen (Online und Live) und der Studie für das Nutzervertrauen sowie die einzelnen Vertrauensdimensionen im Detail untersucht und verglichen.

Nutzervertrauen Das *Nutzervertrauen* in der Live-Umfrage wurde in den meisten Situationen ähnlich bewertet wie in der Onlineumfrage. Allerdings stellte sich heraus, dass die durchschnittlichen Bewertungen des wahrgenommenen Vertrauens in der Live-Umfrage (⌀ 3,91; Bewertung auf der 5er-Likert-Skala) tendenziell höher ausfielen, als in der Onlineumfrage (⌀ 3,16). Dieser Effekt konnte bereits in einer früheren Arbeit beobachtet werden [Wissner et al., 2014]. Es ist anzunehmen, dass Nutzer zu einem System, mit dem sie reale Erfahrungen sammeln konnten, mehr Vertrauen aufbauen als zu einem System, das ihnen nur textuell beschrieben wurde. Dies mag u.a. daran liegen, dass ein Text nicht die kompletten Eindrücke und Emotionen transportieren kann.

Tabelle 6.4: Untersucht: Präferenz - Signifikante Ergebnisse des ANOVA-Tests mit Messwiederholung und des Bonferroni-Post-Hoc-Tests (Abkürzungen: M = Mittelwert; SA = Standardabweichung; K = Kollege; Nichts = Keine Systemreaktion; Auto = Automatisches Ein/Ausschalten; ESP = Empfehlung auf Smartphone; EB = Empfehlung auf Bildschirm)

Gerät	Situation	Signifikanzen (A < B)	Bewertungen			
			M(A)	SA(A)	M(B)	SA(B)
Onlineumfrage - Präferenz						
Licht	Im Raum, Dunkel, K abwesend	ESP < Auto**	2,31	1,06	4,00	1,22
		EB < Auto**	2,00	1,06		
Bildschirm	Aktiv am PC	ESP < Auto***	1,64	0,93	3,86	1,14
		ESP < Nichts***			3,14	1,46
	Abwesend vom PC	ESP < Auto*	1,59	0,94	2,68	1,26
	Inaktiv am PC	ESP < Nichts***	2,05	1,36	4,09	0,95
		Auto < Nichts*	2,91	1,16		
	Abwesend	ESP < Auto**	2,55	1,50	4,09	1,20
Live-Umfrage - Präferenz						
Licht	Im Raum, Dunkel, K abwesend	Nichts < Auto*	2,40	1,28	4,60	0,92
		EB < Auto*	2,60	0,92		
		ESP < Auto**	2,30	1,10		
	Hell, K abwesend	Nichts < Auto**	1,70	0,78	4,10	0,94
		Nichts < EB***			4,50	0,67
		ESP < EB**	2,70	1,19		
	Dunkel, K anwesend	Nichts < EB*	2,30	1,35	4,50	0,67
		ESP < EB**	2,30	1,10		
Bildschirm	Aktiv am PC	Nichts < Auto**	2,80	1,08	4,80	0,40
		ESP < Auto***	1,80	0,87		
	Abwesend vom PC	ESP < Auto*	2,00	1,26	3,90	0,94
		ESP < Nichts*			4,10	0,94
	Inaktiv am PC	Nichts < Auto**	2,70	1,27	4,80	0,40
		ESP < Auto**	2,50	1,20		
	Abwesend	ESP < Auto**	3,30	1,00	4,90	0,30
		Nichts < Auto***	1,70	1,00		
(*signifikant mit p<0,05; **signifikant mit p<0,01; ***signifikant mit p<0,001)						

Die Teilnehmer der Live-Umfrage hielten es für vertrauenswürdiger, wenn das System selbstständig agierte (⌀ 4,18), als wenn es gar nicht tätig wurde (⌀ 3,28). Der Unterschied war allerdings nicht signifikant. Insgesamt erreicht nur „Nichts tun“ keine durchschnittliche Bewertung über 4,0. Da die Systemreaktionen einer vorangegangenen Studie [Wissner et al., 2014] deutlich unterschiedlicher ausgefallen waren, überraschte dieses Resultat. Allerdings bargen die ausgewählten Aktionen in der damaligen Studie teilweise ernsthafte Risiken für private Daten der Nutzer. Dies hatte einen starken Einfluss auf die wahrgenommene Vertrauenswürdigkeit der Aktionen.

Ein Vergleich des wahrgenommenen Vertrauens und der Präferenzen der Teilnehmer zeigte übereinstimmende Resultate für die beiden Kriterien. Autonome Handlungen wurden häufiger vorgezogen (⌀ 4,46) als nichts zu tun (⌀ 2,47) oder über das Smartphone (⌀ 2,59) oder den Bildschirm (⌀ 3,18) Empfehlungen auszusprechen. Siehe auch Tabelle 6.4.

Transparenz In beiden Umfragen sowie in der Studie wurden die meisten der Systemaktionen als *transparent* eingestuft. Die Bewertungen lagen um 4,0 und höher. Die Systemaktionen erschienen den Teilnehmern in den jeweiligen Situationen also plausibel. „Nichts tun“ erhielt jedoch in 75% (Onlineumfrage) und 88% (Live-Umfrage) der Fälle nur mittelmäßige Bewertungen zwischen ca. 2,5 und 3,5. Teilweise wurde „Nichts tun“ sogar signifikant weniger transparent wahrgenommen als das automatische Regeln der Geräte, siehe Tabelle 6.5. Ein möglicher Grund hierfür ist, dass Nutzer proaktives Verhalten, bei dem tatsächlich eine Aktion sichtbar ist, einfacher erkennen und verstehen als „proaktives“ Nichtstun.

Durch die konsistenten Bewertungen über alle Situationen hinweg konnte kein Zusammenhang zu den Präferenzen der Nutzer hergestellt werden.

Kontrollierbarkeit Wie sich in der Onlineumfrage und der Studie bereits gezeigt hatte, führten autonome Systemreaktionen auch in der Live-Umfrage zu niedrigeren Bewertungen der *Kontrollierbarkeit*. Der Durchschnitt lag für alle Situationen unter 3,0. In Tabelle 6.6 ist zu sehen, dass automatische Handlungen des Systems in vielen der Situationen - sowohl in der Onlineumfrage als auch in der Live-Umfrage - signifikant schlechter abschnitten als die anderen Systemaktionen. Besonders gut wurde die Kontrollierbarkeit für Systemaktionen bewertet, die über das Smartphone oder den Bildschirm Empfehlungen präsentierten.

Wie die Analyse der Präferenzen, siehe Tabelle 6.4, zeigte, hatte eine schlechtere Kontrollierbarkeit allerdings keinerlei Einfluss auf die Vorlieben der Nutzer. Obwohl autonome Systemaktionen zu einer geringeren Nutzerkontrolle führten, wurden sie in den meisten Situationen, in denen eine Reaktion erwartet wurde, den anderen Option vorgezogen. Im Gegensatz dazu wurden Empfehlungen auf dem Smartphone als sehr kontrollierbar eingeschätzt. Bei der Angabe der Präferenzen schloss diese Aktion beim Großteil der Situationen mit durchschnittlichen Bewertungen von 3,0 oder weniger nur mittelmäßig ab, siehe Tabelle 6.4.

Tabelle 6.5: Untersuchte Vertrauensdimension: Transparenz - Signifikante Ergebnisse des ANOVA-Tests mit Messwiederholung und des Bonferroni-Post-Hoc-Tests (Abkürzungen: M = Mittelwert; SA = Standardabweichung; K = Kollege; Nichts = Keine Systemreaktion; Auto = Automatisches Ein/Ausschalten; ESP = Empfehlung auf Smartphone; EB = Empfehlung auf Bildschirm)

Gerät	Situation	Signifikanzen (A < B)	Bewertungen			
			M(A)	SA(A)	M(B)	SA(B)
Onlineumfrage - Vertrauensdimension: Transparenz						
Licht	Im Raum, Dunkel, K abwesend	Nichts < Auto*	2,75	1,71	4,31	1,16
	Abwesend, Dunkel, K abwesend	Nichts < Auto*	2,44	1,46	4,00	1,54
Bild-schirm	Aktiv am PC	Nichts < Auto**	2,82	1,53	4,18	1,37
		ESP < Auto*	3,23	1,62		
	Abwesend	Nichts < Auto**	3,18	1,59	4,36	1,15
		ESP < Auto*	3,64	1,52		
(*signifikant mit p<0,05; **signifikant mit p<0,01; ***signifikant mit p<0,001)						

Nutzungskomfort Der Nutzungskomfort autonomer Systemaktionen wurde in beiden Umfragen am besten bewertet. In der Live-Umfrage lagen die durchschnittlichen Bewertungen in den einzelnen Situationen über 4,0 und teilweise sogar höher als 4,5. Die Systemaktionen „Nichts tun“ und speziell „Empfehlung über Smartphone“ schlossen häufig signifikant schlechter ab, siehe Tabelle 6.7. Eine Empfehlung über das Smartphone zu empfangen, erreichte in den meisten Situationen nur mittlere Bewertungen unter 3,0. In einzelnen Fällen lag der Durchschnitt sogar unter 2,0. Damit wurde dieser Aktion der schlechteste Nutzungskomfort zugesprochen.

Betrachtet man die Erkenntnisse über die Präferenzen der Nutzer, so zeigt sich, dass Systemreaktionen, die einen hohen Nutzungskomfort bieten, beliebter sind.

Diskussion Die Ergebnisse der Live-Umfrage lassen darauf schließen, dass die Nutzer komfortablere Systemaktionen mehr kontrollierbareren Aktionen vorzogen. Diese Erkenntnis lässt sich auch durch Aussagen der Nutzer während der Studie und der Live-Umfrage untermauern. Die Teilnehmer begrüßten es, wenn sie teilweise über Empfehlungen die Reaktionen des Systems beeinflussen konnten. Allerdings hielten sie Benachrichtigungen auf ihr Smartphone, wenn überhaupt, nur beim Betreten und Verlassen des Büros für sinnvoll. In allen anderen Situationen nahmen sie diese Form der Benachrichtigung als unbequem und lästig oder sogar aufdringlich war. Häufig war das Smartphone außer Reichweite und die Arbeit musste jedes Mal unterbrochen werden, um die eingegangene Nachricht lesen zu können.

Tabelle 6.6: Untersuchte Vertrauensdimension: Kontrollierbarkeit - Signifikante Ergebnisse des ANOVA-Tests mit Messwiederholung und des Bonferroni-Post-Hoc-Tests (Abkürzungen: M = Mittelwert; SA = Standardabweichung; K = Kollege; Nichts = Keine Systemreaktion; Auto = Automatisches Ein/Ausschalten; ESP = Empfehlung auf Smartphone; EB = Empfehlung auf Bildschirm)

Gerät	Situation	Signifikanzen (A < B)	Bewertungen			
			M(A)	SA(A)	M(B)	SA(B)
Onlineumfrage - Vertrauensdimension: Kontrollierbarkeit						
Licht	Im Raum, Dunkel	Auto < EB**	2,44	1,37	3,50	1,37
		Auto < ESP**			4,00	1,37
	Hell, K abwesend	Auto < Nichts*	1,88	1,17	3,44	1,58
		Auto < ESP**			3,88	1,36
		Auto < EB**			4,00	1,22
	Dunkel, K im Raum	Auto < EB**	1,88	1,05	3,88	1,41
		Auto < ESP**			4,00	1,27
	Abwesend, Dunkel, K abwesend	Nichts < ESP*	2,50	1,54	4,19	1,13
		Auto < ESP*	2,63	1,49		
Bildschirm	Abwesend vom PC	Auto < ESP**	2,27	1,17	3,73	1,39
		Auto < Nichts***			3,59	1,40
	Inaktiv am PC	Auto < Nichts**	2,23	1,04	3,45	1,53
		Auto < ESP**			3,64	1,30
Live-Umfrage - Vertrauensdimension: Kontrollierbarkeit						
	Im Raum, Dunkel, K abwesend	Auto < EB**	2,40	1,36	4,50	0,67
		Auto < ESP**			4,70	0,46
Licht	Hell, K abwesend	Auto < ESP**	2,30	1,19	4,60	0,49
		Auto < EB**			4,80	0,40
	Dunkel, K anwesend	Auto < ESP**	2,40	1,20	4,80	0,40
		Auto < EB**			4,90	0,30
	Abwesend, Dunkel, K abwesend	Auto < ESP**	2,30	1,19	4,80	0,40
Bildschirm	Aktiv am PC	Auto < ESP**	2,80	1,17	4,50	0,50
	Abwesend vom PC	Auto < ESP***	2,30	1,19	4,70	0,46
	Inaktiv am PC	Nichts < ESP*	3,30	1,55	4,70	0,46
		Auto < ESP***	2,30	0,78		
	Abwesend	Auto < ESP**	2,60	1,28	4,80	0,40
(*signifikant mit p<0,05; **signifikant mit p<0,01; ***signifikant mit p<0,001)						

In der Onlineumfrage, deren Daten zur Initialisierung der UTMs genutzt wurden, war dieses Problem weniger stark ins Gewicht gefallen. Es trat durch die verbale Beschreibung der Szenarien womöglich weniger in Erscheinung. In einer früheren Arbeit [Wissner et al., 2014], bei der die Gefährdung der privaten Daten im Vordergrund stand, trat dieser Effekt nicht auf. Die Bedrohung der eigenen Privatsphäre wirkte auch in der textbasierten Beschreibung der Situationen so stark auf die Teilnehmer der Befragung ein, dass die in der damaligen Onlineumfrage gewonnen Daten sehr gut zur Initialisierung des UTM geeignet waren. Im Falle der SavER-Anwendung war auf die online gesammelten Daten allerdings nur begrenzt Verlass.

In zukünftigen Arbeiten sollten deshalb zwei Maßnahmen in Erwägung gezogen werden: (1) Untersuchungen, ob manche Vertrauensdimensionen die Präferenzen und das Vertrauen der Nutzer stärker beeinflussen als andere. (2) Initialisierung des UTM mit einer ausreichenden Menge an Daten, die „live“, d.h. während die Nutzer die Situationen tatsächlich am eigenen Körper erleben, gewonnen werden.

Eine weitere Verbesserung der Initialisierung des UTM könnte auch durch eine Anpassung der Datensammlung erreicht werden. Um die Teilnehmer der Befragungen mit realistischen Szenarien zu konfrontieren, wurden die einzelnen Aufgaben und Situationen in eine kohärente Geschichte eingegliedert, die einen kompletten Arbeitstag umfasste. Der Nachteil dieses Vorgehens ist allerdings, dass alle Teilnehmer die gleiche Sequenz durchliefen. Da auch die Evaluation des Systems mit dem selben Ablauf erfolgte, ist ein Overfitting des UTM nicht auszuschließen. Zukünftige Datensammlungen sollten deshalb mehr und längere Szenarien beinhalten, die in wechselnden Reihenfolgen durchlaufen werden können.

6.4 Zusammenfassung

Die Forschungsfrage in diesem Kapitel lautete: Ist es möglich das Vertrauen der Nutzer gegenüber einem beratenden Empfehlungssystem wie SavER und seinen Entscheidungen so zu modellieren, dass das System darauf aufbauend autonome Entscheidungen über die Angemessenheit verschiedener Systemaktionen fällen kann? Im Zusammenhang damit sollten entscheidende Einflussfaktoren für das Nutzervertrauen identifiziert und die Einschätzung bestimmter Systemaktionen durch die Nutzer ermittelt werden.

Mit dem User Trust Model (UTM) wurde ein auf Bayes'schen Netzen basierender Ansatz vorgestellt, der das Vertrauen der Nutzer in ein System und seine Handlungen abschätzen kann. Am Beispiel des SavER-Szenarios wurde gezeigt, dass die grundlegende und zum Teil vorinitialisierte Struktur des UTM gut in beratende Empfehlungssysteme integriert werden kann.

Die Ergebnisse einer Evaluation und einer Live-Umfrage zeigten, dass das SavER-System durch das UTM tatsächlich dazu befähigt wurde, abhängig von der vorliegenden Situation angemessene bzw. vertrauenswürdige Systemreaktionen auszuwählen. Die Studienteilnehmer hatten in den meisten Situationen großes Vertrauen in die Entscheidungen des Systems und akzeptierten das System in dieser Form.

Tabelle 6.7: Untersuchte Vertrauensdimension: Nutzungskomfort - Signifikante Ergebnisse des ANOVA-Tests mit Messwiederholung und des Bonferroni-Post-Hoc-Tests (Abkürzungen: M = Mittelwert; SA = Standardabweichung; K = Kollege; Nichts = Keine Systemreaktion; Auto = Automatisches Ein/Ausschalten; ESP = Empfehlung auf Smartphone; EB = Empfehlung auf Bildschirm)

Gerät	Situation	Signifikanzen (A < B)	Bewertungen M(A)	SA(A)	M(B)	SA(B)
Onlineumfrage - Vertrauensdimension: Nutzungskomfort						
Licht	Im Raum, Dunkel, K abwesend	ESP < Auto*	2,50	1,37	3,88	1,22
		EB < Auto**	2,25	1,09		
		Nichts < Auto**	2,06	0,90		
Bildschirm	Aktiv am PC	ESP < Nichts*	1,73	0,86	2,64	1,30
		ESP < Auto***			3,86	1,36
		Nichts < Auto**	2,64	1,30	3,86	1,36
	Abwesend vom PC	ESP < Auto*	1,86	0,97	2,64	1,33
	Inaktiv am PC	ESP < Auto*	2,05	1,11	2,91	1,16
		ESP < Nichts*			3,36	1,26
	Abwesend	ESP < Auto*	2,64	1,40	3,91	1,41
Live-Umfrage - Vertrauensdimension: Nutzungskomfort						
Licht	Im Raum, Dunkel, K abwesend	ESP < Auto*	2,50	1,12	4,50	0,67
	Hell, K abwesend	Nichts < Auto**	2,40	1,20	4,60	0,66
		ESP < Auto*	2,70	0,90		
		Nichts < EB*	2,40	1,20	4,20	0,40
		ESP < EB**	2,70	0,90		
	Dunkel, K anwesend	Nichts < Auto**	2,50	1,12	4,60	0,66
		ESP < Auto*	2,60	1,20		
		Nichts < EB**	2,50	1,12	4,30	0,46
		ESP < EB**	2,60	1,20		
	Abwesend, Dunkel, K abwesend	ESP < Auto*	3,50	1,12	4,80	0,40
		Nichts < Auto**	2,50	1,12		
Bildschirm	Aktiv am PC	Nichts < Auto**	2,50	1,20	4,70	0,46
		ESP < Auto***	2,00	0,63		
	Abwesend vom PC	ESP < Auto*	1,90	1,22	3,90	1,14
	Inaktiv am PC	ESP < Auto*	2,50	1,20	4,30	1,00
(*signifikant mit p<0,05; **signifikant mit p<0,01; ***signifikant mit p<0,001)						

Mit Fokus auf die einzelnen Dimensionen von Nutzervertrauen zeigte sich, dass die Nutzer häufig zu Gunsten eines höheren Nutzungskomforts weniger vertrauenswürdige und kontrollierbare Aktionen bevorzugten. Dieses Ergebnis deckt sich mit den Erkenntnissen von Barkhuus und Dey [Barkhuus und Dey, 2003]. Sie fanden heraus, dass Nutzer dazu bereit sind einen Teil ihrer Kontrolle über (semi-)autonome Systeme abzugeben, wenn sie dafür andere Vorteile wie einen besseren Nutzungskomfort erhalten. Bei einem zu drastischen Kontrollverlust oder zu geringen Vorteilen, kann es aber zu Frustrationen und einer Ablehnung des jeweiligen Systems kommen.

Ein wichtiger Punkt für zukünftige Arbeiten ist eine stärkere Personalisierung der Entscheidungsfindung. Im evaluierten SavER-System reflektierten die Entscheidungen des Systems die Meinungen der Personen, deren Daten zur Initialisierung des Systems genutzt wurden. Der Einfluss des individuellen *Nutzernaturells* auf das Nutzervertrauen und auch die individuelle Präferenz für einzelne Vertrauensdimensionen sollten deswegen künftig näher untersucht und bei der Initialisierung des Systems für die einzelnen Nutzer berücksichtigt werden. Eine Person, die generell skeptisch gegenüber technischen Systemen ist, könnte mehr Wert auf ein hohes Maß an Kontrolle legen. Eine Person mit großer Kompetenz und viel Vertrauen in technische Systeme könnte stattdessen einen hohen Nutzungskomfort bevorzugen. Ein vielversprechender Ansatz für eine Umsetzung wäre eine multidimensionale Kategorisierung der Nutzer nach dem Beispiel von Knijnenburg und Kollegen [Knijnenburg et al., 2013], die zum Beispiel die Demographie, die Persönlichkeit, die Erfahrung und das Naturell der Nutzer berücksichtigten. Die Daten für erste Untersuchungen in diese Richtung wären im Kontext des SavER-Systems durch den einleitenden Fragebogen der Live-Umfrage bereits vorhanden. Dort wurden die Teilnehmer zum einen nach ihrer Meinung und ihren Gewohnheiten im Bezug auf nachhaltiges Verhalten befragt. Zum anderen wurde aber auch ihr generelles Vertrauen gegenüber anderen Personen und technischen Systemen erfragt.

Ein langfristiges Ziel sollte ein dynamisches UTM sein, dass möglichst ohne aufwendiges Training und eine teure Datensammlung implizit und dynamisch während der Laufzeit aus dem Verhalten seiner Nutzer auf ihre Präferenzen und ihr Vertrauen gegenüber dem System schließen kann. Zum Beispiel könnte auf ein erhöhtes Vertrauen und eine gestiegene Relevanz automatischer Systemreaktionen geschlossen werden, wenn eine Person dem System gegenüber Vertrauen zeigt und Kontrolle abgibt [Lee und See, 2004]. Durch ein dynamisches UTM könnte auch berücksichtigt werden, dass das Vertrauen in ein System und seine Entscheidungen nicht nur von der aktuellen Situation abhängt, sondern von den Erfahrungen, die eine Person über einen längeren Zeitraum mit dem System gesammelt hat. Eine einmalige Fehlentscheidung des Systems mit geringen negativen Auswirkungen wird das Nutzervertrauen weniger beeinflussen, als wiederholte oder schwerwiegendere Fehler. Technisch wäre diese Weiterentwicklung des UTM durch einer Erweiterung des Bayes'schen Netzes zu einem dynamischen Bayes'schen Netz möglich.

Ein letzter wichtiger Aspekt, der in zukünftigen Arbeiten angegangen werden müsste, ist die Berücksichtigung von mehr als einer Person im Entscheidungsprozess.

In der bisherigen Variante des SavER-Systems wurde lediglich eine Person als aktiver Nutzer berücksichtigt und die restlichen Personen im Umfeld als sozialer Kontext angesehen. In den Evaluationen wunderten sich allerdings einige der Teilnehmer, warum nur sie alleine die Entscheidungen über den Status des Lichts treffen sollten. Das Vertrauen aller betroffener Nutzer könnte durch individuelle UTMs für alle Personen in die Entscheidungsfindung miteinbezogen werden. Für die Kombination der verschiedenen Nutzerpräferenzen könnten Ansätze für Gruppenempfehlungssysteme genutzt werden [Masthoff, 2011].

7 Schluss

Empfehlungssysteme haben seit den 90er-Jahren vor allem im E-Commerce-Bereich stark an Bedeutung gewonnen. Sie unterstützen in Zeiten der Daten- und Informationsflut die menschliche Entscheidungsfindung beim Kauf von Produkten oder bei der Suche nach Wohnungen, Restaurants oder Sehenswürdigkeiten. Der Einsatz von Empfehlungssystemen als persönlicher Assistent, der die Nutzer bei alltäglichen Herausforderungen wie der Förderung ihrer Gesundheit oder der Optimierung ihres Energieverbrauchs unterstützt, wurde bisher nur in einzelnen Forschungsarbeiten in Betracht gezogen. Solch beratende Empfehlungssysteme sprechen (proaktiv) domänenspezifische Empfehlungen für Aktivitäten und Maßnahmen aus und versuchen dadurch u.a. die Kompetenz, das Selbstbewusstsein und die Motivation der Nutzer im jeweiligen Bereich zu steigern.

In dieser Dissertation wurde untersucht, wie beratende Empfehlungssysteme Wissen über menschliche Werte und Verhaltensweisen nutzen können, um durch situative und personalisierte Entscheidungen während des Empfehlungsprozesses die UX-Faktoren Überzeugungskraft, Nutzerakzeptanz und Nutzervertrauen positiv beeinflussen zu können. Die in dieser Arbeit betrachteten Schritte des Empfehlungsprozesses umfassten die Empfehlungsauswahl, die Generierung von Empfehlungstexten und das proaktive Ausführen von Empfehlungen durch das System selbst.

Für jeden der genannten Schritte wurden geeignete psychologische und sozialwissenschaftliche Theorien und Modelle identifiziert und Ansätze vorgestellt, wie diese Theorien und Modelle in die Entscheidungsfindung der Empfehlungssysteme integriert werden können. Als beispielhafte Anwendungsszenarien dienten die Förderung des Wohlbefindens alleinstehender Senioren (CARE) und die Förderung energiesparenden Verhaltens (SavER). Innerhalb dieser beiden Anwendungsszenarien wurden die vorgestellten Ansätze prototypisch umgesetzt und evaluiert.

Im Folgenden werden die wissenschaftlichen Beiträge dieser Arbeit zusammengefasst und diskutiert. Außerdem werden Forschungsthemen vorgestellt, die auf dieser Dissertation aufbauen und die entstandenen Ansätze ergänzen könnten.

7.1 Wissenschaftliche Beiträge

Die wissenschaftlichen Beiträge dieser Dissertation können anhand der betrachteten Schritte innerhalb des Empfehlungsprozesses in einem beratenden Empfehlungssystem kategorisiert werden: Empfehlungsauswahl, Generierung von Empfehlungstexten und proaktives Ausführen von Empfehlungen

Empfehlungsauswahl In Empfehlungssystemen, die eine kollaborative Filterung durchführen, treten aufgrund mangelnder Nutzerbewertungen häufig Probleme wie das Sparsity-Problem oder das New-User-Problem auf. Daher werden seit längerem Ansätze erforscht, die die Qualität der kollaborativen Empfehlungsauswahl in Cold-Start-Szenarien verbessern können.

Im Hinblick auf beratende Empfehlungssysteme wurde in dieser Arbeit ein auf sozialwissenschaftlichen Theorien basierender Ansatz untersucht, der Nutzermodelle berücksichtigt, die domänenspezifisches Wissen über die Werte, Fähigkeiten und Verhaltensweisen der Nutzer enthalten, siehe Kapitel 4.3. Die Annahme war, dass Nutzer, die sich hinsichtlich solcher theoriebasierter Nutzermodelle ähneln, auch durch ähnliche Aktivitäten und Maßnahmen gefördert werden können.

Für die Integration theoriebasierter Nutzermodelle in das klassische kollaborative Filterverfahren wurden drei Varianten vorgestellt. In der einfachsten Variante wird die Ähnlichkeit von Nutzern lediglich anhand des theoriebasierten Nutzermodells bewertet. Die beiden anderen Varianten stellen hybride Formen der Filterung dar und kombinieren den bewertungsbasierten Ansatz mit dem theoriebasierten Ansatz. Im einen Fall werden die Ergebnisse einer rein bewertungsbasierten und einer rein theoriebasierten Ähnlichkeitsberechnung linear kombiniert. Im anderen Fall werden durch den theoriebasierten Ansatz zusätzliche künstliche Bewertungen für die anschließende bewertungsbasierte Filterung erzeugt (Merkmalserweiterung).

Zur Evaluierung der Filterverfahren wurde zunächst nach geeigneten Modellen recherchiert, die für die Empfehlungsauswahl in den Anwendungsszenarien CARE und SavER geeignet waren, siehe Kapitel 4.2. Bei CARE lag es nahe, dass u.a. das aktuelle körperliche, geistige und mentale Wohlbefinden der Nutzer ein entscheidender Faktor für die Einschätzung spezifischer Empfehlungen ist. Im SavER-System werden die Nutzer dagegen durch ihre allgemeine Einstellung zum Thema Energiesparen sowie ihre Möglichkeiten und ihr bisheriges Verhalten beeinflusst.

Wie die Evaluationen in beiden Anwendungsszenarien, die in Kapitel 4.3 beschrieben wurden, zeigten, konnten durch die Integration der theoriebasierten Nutzermodelle in Cold-Start-Szenarien sowohl für die Vorhersage von Bewertungen als auch für die Klassifikation der möglichen Empfehlungen hinsichtlich ihrer Relevanz für die Nutzer signifikante Verbesserungen erzielt werden. Die beste Qualität der Empfehlungsauswahl konnte durch die beiden hybriden Filterverfahren (lineare Kombination, Merkmalserweiterung) erzielt werden.

Generierung von Empfehlungstexten Eines der wichtigsten Postulate persuasiver Systeme ist, dass Informationstechnologie niemals neutral ist [Oinas-Kukkonen und Harjumaa, 2009]. Sobald eine Interaktion zwischen Nutzer und System stattfindet, werden die Meinung und das Verhalten der Person beeinflusst. Des Weiteren ist bekannt, dass Nutzer sprach- und textbasierten Systemen basierend auf der Wortwahl automatisch eine Persönlichkeit zuordnen. Dies beeinflusst wiederum die Beziehung zwischen Nutzer und System [Reeves und Nass, 1998].

Ein Ziel dieser Dissertation war es, dieses Wissen zu nutzen, um die User Experience (Nutzervertrauen, Nutzerakzeptanz, Überzeugungskraft) beratender Empfehlungssystemen zu beeinflussen. Hierfür wurden drei Faktoren erforscht, die laut Literatur [Mairesse, 2008, Marcu, 1996] die Wahrnehmung natürlichsprachlicher Interaktionen beeinflussen können: Kulturbasierte Auswahl von Argumenten, Höflichkeitsstrategien und Persönlichkeitsausprägungen von Systemen

Kulturbasierte Auswahl von Argumenten Wichtige Erkenntnisse der Literaturrecherche in Kapitel 5.1 waren, dass Argumente für Empfehlungen personalisiert werden sollten und dass die Wahrnehmung von Argumenten grundsätzlich kulturabhängig ist. Darauf aufbauend wurde untersucht, ob die Kultur der Teilnehmer tatsächlich einen Einfluss auf die Überzeugungskraft von Argumenten für Energiesparempfehlungen hat und ob Hofstedes Kulturmodell [Hofstede, 2001] dafür geeignet ist, eine kulturbasierte Argumentauswahl zu realisieren.

In einer Online-Studie mit Teilnehmern aus 15 Ländern wurden sowohl von den Teilnehmern für ihre Kulturgruppe formulierte Argumente als auch die Bewertungen der Teilnehmer für vorgegebene Argumente analysiert. Die Ergebnisse zeigten, dass eine personalisierte und kulturabhängige Auswahl von Argumenten sich durchaus positiv auf die Überzeugungskraft von Argumenten auswirken kann. Allerdings galten Themen wie Geld, Energiesparen und Umweltschutz interkulturell als wichtig und überzeugend. Des Weiteren wurde deutlich, dass eine Fokussierung auf Hofstedes Kulturdimensionen für die Argumentauswahl nicht ausreichend ist, da eine Einschätzung der Nutzer nur anhand ihres Herkunftslandes zu ungenau ist. Um eine gute Personalisierung von Argumenten erreichen zu können, ist eine stärkere Personalisierung der Argumentauswahl, die u.a. auf der Energiekultur der Nutzer basieren könnte, von Nöten. Das Aufgreifen konkreter und aktueller Informationen wie eingesparter Kosten, aktueller Spritpreise oder des aktuellen Wetters könnten außerdem langfristig die Beachtung der Argumente durch die Nutzer fördern.

Höflichkeitsstrategien in Empfehlungstexten In beratenden Empfehlungssystemen besteht die Gefahr, dass die ausgesprochenen Empfehlungen bei ihren Nutzern ein Gefühl von Scham oder Bevormundung hervorrufen, da sie die Nutzer (indirekt) auf ihre Schwächen oder ein möglicherweise falsches Verhalten hinweisen. Für die Formulierung der Empfehlungstexte muss daher situativ eine Balance zwischen Höflichkeit und Überzeugungskraft gefunden werden. Aus diesem Grund wurde in Kapitel 5.2 untersucht, ob die Höflichkeitsstrategien von Brown und Levinson [Brown und Levinson, 1987] dazu geeignet sind, die wahrgenommene Höflichkeit und Überzeugungskraft von Empfehlungen zu steuern.

Sowohl in einer textbasierten Evaluation mit jüngeren Teilnehmern als auch in einer Wizard-of-Oz-Studie im CARE-Szenario mit älteren Menschen konnte bestätigt werden, dass die auf den untersuchten Strategien basierenden Formulierungen als unterschiedlich höflich wahrgenommen wurden. Die zum ersten Mal untersuchte Überzeugungskraft von Höflichkeitsstrategien in Empfehlungstexten unterschied sich jedoch nur in der Evaluation mit den jüngeren Nutzern. Dennoch konnten anhand der Ergebnisse beider Evaluationen Strategien für den Einsatz verschiedener Formulierungen in beratenden Empfehlungssystemen abgeleitet werden. Formulierungen als Fragen oder als Ziele des Systems, die zwar als höflich, aber nicht als besonders überzeugend bewertet wurden, könnten eingesetzt werden, wenn die Beziehung zwischen Nutzer und System gepflegt werden soll. Falls jedoch eine gute Balance aus Höflichkeit und Überzeugungskraft gefragt ist, könnten Empfehlungen

als gemeinsames Ziel oder als Anfrage bzw. Bitte formuliert werden, da diese Varianten als höflich und überzeugend eingeschätzt werden. Direkte Kommandos sind dagegen dann geeignet, wenn die Überzeugungskraft im Fokus steht und eine reduzierte Höflichkeit auf die Dringlichkeit einer Empfehlung hinweisen soll.

Persönlichkeitsausprägungen von Formulierungen Menschen ordnen einem System, das per Sprache oder Text mit ihnen kommuniziert, automatisch eine Persönlichkeit zu. Zusätzlich agieren sie mit diesen Systemen auch auf eine ähnlich soziale Weise, wie sie es mit anderen Menschen tun. In verschiedensten Forschungsarbeiten wurde deshalb bereits untersucht, wie gezielte Formulierungen von Systemausgaben genutzt werden können, um die individuelle Wahrnehmung des jeweiligen Systems zu beeinflussen. In Kapitel 5.3 wurde nun zum ersten Mal untersucht, ob die bisherigen Erkenntnisse der Forschung in beratenden Empfehlungssysteme aufgegriffen werden können, um auf die wahrgenommene Überzeugungskraft und die wahrgenommene Vertrauenswürdigkeit der Systeme einwirken zu können.

Mittels einer prototypischen Umsetzung einer Sprachausgabe für ein CARE-System mit Hilfe des Frameworks PERSONAGE von Mairesse und Walker [Mairesse und Walker, 2007] wurde gezeigt, wie Empfehlungstexten eine gewünschte Persönlichkeitsausprägung verliehen werden kann.

In einer anschließenden Studie wurden drei Strategien zur Adaption der Systempersönlichkeit an die individuellen Nutzer untersucht. Die erste Strategie nutze eine neutrale Persönlichkeit, die zweite Strategie spiegelte die Persönlichkeit der Zielperson wider (Mirroring) und die letzte Strategie wählte bewusst eine Persönlichkeit, die der Persönlichkeit der Zielperson entgegensteht (Mismatching).

Als vielversprechend stellte sich die Strategie des Mirroring heraus, da sie die Vertrauenswürdigkeit der Empfehlungstexte signifikant verbessern konnte. Eine Ursache dafür, dass keine weiteren signifikanten Unterschiede festgestellt werden konnten, könnte sein, dass die vom System generierten Formulierungen sich in der Studie großteils nur in feinen Details unterschieden. Deutlichere Unterschiede hinsichtlich klassischer sprachlicher Merkmale für unterschiedliche Persönlichkeitsausprägungen wie die Satzlänge, die Polarität der Sätze oder die Häufigkeit der Nachfragen könnten die Persönlichkeit eines Systems stärker hervorheben und somit zu größeren Unterschieden bei der Einschätzung der Vertrauenswürdigkeit und Überzeugungskraft der Empfehlungstexte führen.

Autonome Ausführung von Empfehlungen Selbst, wenn ein beratendes Empfehlungssystem immer die richtigen Empfehlungen auswählen und mit einem geeigneten Empfehlungstext präsentierten sollte, können vermehrt oder in ungünstigen Situationen auftretende Empfehlungen aufdringlich, störend oder lästig erscheinen. Des weiteren könnten die Nutzer auch das Gefühl bekommen nicht mehr ausreichend Kontrolle über das Empfehlungssystem zu haben.

Als Lösung dieses Problem wurde in Kapitel 6 der Ansatz untersucht, dass einem assistierenden Empfehlungssystem für einzelne Maßnahmen wie dem Ausschalten

von Geräten ein gewisses Maß an Autonomie zugestanden wird. So könnten die Systeme in manchen Situationen kleinere Missstände selbstständig beheben, ohne dass die Nutzer ihre aktuellen Tätigkeiten unterbrechen und selbst aktiv werden müssten. Ein wichtiges Entscheidungskriterium für die Angemessenheit autonomen Systemverhaltens stellt das Vertrauen der Nutzer gegenüber dem System dar.

In dieser Dissertation wurde mit dem User Trust Model (UTM) ein Modell vorgestellt, das es intelligenten Systemen zum ersten Mal ermöglicht, das Nutzervertrauen gegenüber sich und seinen Aktionen einzuschätzen und basierend darauf autonom die Systemaktion auszuwählen, die das größte Nutzervertrauen zur Folge hat.

Die Evaluation eines Prototypen im SavER-Szenario zeigte, dass das SavER-System durch die Integration des UTM dazu in der Lage war vertrauenswürdige Entscheidungen zu treffen. Auch die Ergebnisse hinsichtlich der Nutzerakzeptanz waren vielversprechend. Etwas überraschend war es dagegen, dass die Systemreaktion, der das größte Vertrauen entgegengebracht wurde, nicht immer auch die präferierte Systemaktion der Nutzer war. Eine Live-Umfrage zur Klärung offener Fragen aus der Evaluation ergab, dass ein Großteil der Nutzer im SavER-Szenario dazu bereit war, zugunsten eines besseren Nutzungskomforts auf Kontrolle über das System zu verzichten. Sie präferierten daher in manchen Situationen weniger vertrauenswürdige Systemaktionen.

Um solche Effekte noch besser durch das UTM abbilden zu können, müsste dieses noch stärker personalisiert werden. In dieser Dissertation wurde das Modell mit den Daten vieler verschiedener Menschen initialisiert, so dass die Entscheidungen des Systems eher auf dem generellen Konsens der Befragten beruhten als auf den individuellen Einstellungen und Werten.

7.2 Fortführende Arbeiten

Diese Dissertation legt den Grundstein für den Entwurf und die Entwicklung beratender Empfehlungssysteme. Es wird jedoch weiterführende Forschung benötigt, um die User Experience assistierender Empfehlungssysteme und damit auch ihre Erfolgschancen weiter verbessern zu können.

Dynamische Nutzermodelle Die Performanz der Filtertechnologien mit theoriebasierten Nutzermodellen wurde, wie in Kapitel 4.3.3 beschrieben, mit einem statischen Datensatz evaluiert. Auf längere Sicht ist eines der Hauptziele beratender Empfehlungssysteme jedoch, die Einstellung bzw. das Verhalten der Nutzer zu verändern. In diesem Fall müssen sich die veränderten Werte, Fähigkeiten und Präferenzen der Nutzer sich auch in ihren Nutzermodellen widerspiegeln. Die Nutzermodelle müssten demzufolge dynamisch angepasst werden können. Dabei sollte allerdings darauf geachtet werden, dass Änderungen der Präferenzen sowohl kurzfristiger als auch langfristiger Art sein können. Zum Beispiel hat die aktuelle Tagesform älterer Menschen einen kurzfristigen Einfluss auf ihre körperlichen oder mentalen Fähigkeiten. Eine regelmäßige Durchführung förderlicher Aktivitäten kann jedoch zu einer

andauernden Verbesserung dieser Fähigkeiten führen. Zukünftige Arbeiten sollten sich damit beschäftigen, wie Nutzermodelle durch explizite Befragungen, aber auch durch implizite Analysen des Nutzerverhaltens aktuell gehalten werden können. Außerdem sollten die Auswirkungen dynamischer Nutzermodelle auf die Qualität der Empfehlungsauswahl untersucht werden.

Dialoge In dieser Dissertation bestand die Interaktion zwischen System und Nutzer aus der Präsentation einer oder mehrerer Empfehlungen und der anschließenden Annahme oder Ablehnung dieser Empfehlung(en) durch die Nutzer. Eine wertvolle Erweiterung des bisherigen Interaktionskonzeptes wäre die Einführung von Dialogen. Durch Dialoge könnten Faktoren wie die Höflichkeit und die Persönlichkeit eines beratenden Empfehlungssystems noch besser herausgestellt und die soziale Bindung zwischen System und Zielperson stärker gefördert werden. Außerdem könnten mehrere Argumente für und gegen die Annahme von Empfehlungen ausgetauscht und dadurch auch die themenspezifische Kompetenz der Nutzer gesteigert werden. Arbeiten aus dem Bereich der dialogorientierten (engl. conversational) [Carenini et al., 2003, Ikemoto et al., 2018, Mahmood und Ricci, 2009] bzw. auf Kritik basierten (engl. critiquing-based) Empfehlungssysteme [Chen und Pu, 2012, McGinty und Reilly, 2011] könnten eine gute Grundlage für die zukünftige Forschung in diese Richtung darstellen.

Interaktionsgeräte Abhängig vom Anwendungsszenario und der jeweiligen Zielgruppe könnte auch das eingesetzte Interaktionsgerät einen Einfluss auf die UX und die Akzeptanz assistierender Empfehlungssysteme haben. In einer ersten Studie, die gegen Ende dieser Dissertation im ForGenderCare Projekt durchgeführt wurde, wurde beispielsweise bereits der Einsatz von Tablet PCs und sozialen Robotern im CARE-Szenario verglichen, siehe Abbildung 7.1 und [Hammer et al., 2017]. Kurz zusammengefasst lässt sich sagen, dass dem Roboter, obwohl lediglich Empfehlungen präsentiert wurden und keine weiteren Interaktionen stattfanden, eine höhere Usability zugesprochen wurde als dem System auf dem Tablet PC. Unter anderem wurden die Komplexität und die Erlernbarkeit des Systems beim Roboter besser bewertet. Allerdings wurden beide Systeme generell positiv aufgenommen, so dass es bzgl. anderer Faktoren wie zum Beispiel der wahrgenommenen Überzeugungskraft nur leichte Tendenzen zu Gunsten des Roboters gab.

Abbildung 7.1: Empfehlungen der Kategorien physisches und emotionales Wohlbefinden präsentiert von einem sozialen Roboter (links) und über einen Tablet PC

In zukünftigen Arbeiten sollten die Ergebnisse dieser Studie nochmals genauer untersucht werden und auch andere Geräte wie Smartphones oder Smartwatches berücksichtigt werden.

Non-verbales Verhalten und Aussehen sozialer Agenten Laut Nomura und Saeki [Nomura und Saeki, 2010] können bereits kleine Unterschiede bei Bewegungen oder Posen sozialer Roboter sowohl die Wahrnehmung des Roboters als auch das anschließende Verhalten der Menschen gegenüber dem Roboter beeinflussen. Setzt man virtuelle Agenten und soziale Roboter als Interaktionsgerät ein, spielen demzufolge neben der sprachlichen Ausformulierung von Empfehlungstexten auch non-verbale Faktoren wie Gestik oder Körperhaltung, aber auch Aussehen, Mimik und Blickverhalten eine große Rolle bei der Wahrnehmung eines beratenden Empfehlungssystems. Das mögliche Potential non-verbalen Verhaltens zur Verbesserung der UX (z.B. Wahrnehmung der Höflichkeit und der Persönlichkeit) assistierender Empfehlungssystem sollte in zukünftigen Arbeiten näher untersucht werden.

Langzeitstudien Die Evaluationen dieser Dissertation beschäftigten sich hauptsächlich mit der unmittelbaren UX in konkreten Szenarien. Es wurde davon ausgegangen, dass einzelne positive Nutzererlebnisse die Grundlage für eine längerfristige Nutzung beratender Empfehlungssysteme und damit auch für Verhaltensänderungen darstellen. Durch länger andauernde Evaluationen mit vollständigen Systemen sollte in Zukunft u.a. evaluiert werden, ob die neuen Verfahren zur Empfehlungsauswahl tatsächliche Verhaltensänderungen bei den Nutzern zur Folge haben. Die in dieser Dissertation verwendeten Evaluationsmetriken gehören in der Erforschung von Empfehlungssystemen zwar zum Standard, sehr gute Ergebnisse bei diesen Metriken sind aber nicht gleichbedeutend mit einer erfolgreichen Überzeugung von den empfohlenen Maßnahmen und Aktivitäten. Jannach et al. [Jannach et al., 2016] geben zu Bedenken, dass die genaue Vorhersage niedriger Bewertungen zum Beispiel kaum Einfluss auf die wahrgenommene Qualität eines Empfehlungssystems hat, da diese Objekte sowieso nicht empfohlen werden sollten. Ähnlich verhält es sich bei

der Einschätzung der Relevanz von Objekten. Während ein nicht empfohlenes, relevantes Objekt keinen direkten Einfluss auf die Wahrnehmung des Systems haben sollte, kann die Empfehlung eines irrelevanten Objekts (abhängig von der Domäne) die Meinung der Nutzer über das Systems stark negativ beeinflussen.

Ethische Fragen Wie bei allen persuasiven Systemen muss auch hinsichtlich der in dieser Dissertation entwickelten und untersuchten Technologien auf ethische Fragen eingegangen werden. Zum einen können diese Technologien missbraucht werden, um Personen dazu zu bringen Dinge zu tun oder zu erwerben, die für sie keinen Nutzen haben oder ihnen womöglich sogar schaden. Zum anderen stellt sich aber auch die Frage, inwiefern es ethisch vertretbar ist, dass ein möglicherweise fehlerhaftes System Empfehlungen für Maßnahmen ausspricht, die das Verhalten und Wohlbefinden der Nutzer beeinflussen sollen. Im CARE-Projekt wurde die rote Linie bei der Einnahme von Medikamenten gezogen. Doch auch augenscheinlich gut gemeinte Empfehlungen für zum Beispiel Spaziergänge oder bestimmte Rezepte können bei fehlenden, falschen oder missinterpretierten Daten negative Auswirkungen auf ihre Empfänger haben. Bevor die in dieser Dissertation untersuchten Technologien also in realen Systemen eingesetzt werden, sollte intensiv auf sich stellende ethische Fragen eingegangen werden.

Kombination mit anderen persuasiven Techniken Empfehlungssysteme sollten nicht als isoliertes Werkzeug betrachtet werden, wenn Menschen ihr Verhalten ändern oder ihr Wohlbefinden steigern wollen. Diese Dissertation zeigte zwar das große Potential beratender Empfehlungssysteme auf. Eine Kombination mit anderen persuasiven Techniken wie personalisiertem Feedback [Gamberini et al., 2012, Froehlich et al., 2009] oder Gamification-Ansätzen [Gustafsson et al., 2010, Hamari et al., 2014] könnte den Einfluss auf die Nutzer aber weiter verstärken und noch bessere Ergebnisse für die UX der Systeme und die persönlichen Ziele der Nutzer ermöglichen.

Literatur

[Aaker, 1997] Aaker, J. L. (1997). Dimensions of Brand Personality. Journal of Marketing Research, 34(3):347–356.

[Aaker und Maheswaran, 1997] Aaker, J. L. und Maheswaran, D. (1997). The effect of cultural orientation on persuasion. Journal of Consumer Research, Gainesville, 24:315–328.

[Aamodt und Plaza, 1994] Aamodt, A. und Plaza, E. (1994). Case-based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Commun., 7(1):39–59.

[Abdul-Rahman und Hailes, 1997] Abdul-Rahman, A. und Hailes, S. (1997). A Distributed Trust Model. In Proceedings of the 1997 Workshop on New Security Paradigms, NSPW '97, Seiten 48–60, New York, NY, USA. ACM.

[Adomavicius und Tuzhilin, 2005] Adomavicius, G. und Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowl. and Data Eng., 17(6):734–749.

[Adomavicius und Tuzhilin, 2011] Adomavicius, G. und Tuzhilin, A. (2011). Context-Aware Recommender Systems. In Ricci, F., Rokach, L., Shapira, B., und Kantor, B. P., Herausgeber, Recommender Systems Handbook, Seiten 217–253. Springer US, Boston, MA.

[Ahn, 2008] Ahn, H. J. (2008). A New Similarity Measure for Collaborative Filtering to Alleviate the New User Cold-starting Problem. Inf. Sci., 178(1):37–51.

[Anders et al., 2013] Anders, G., Steghöfer, J.-P., Klejnowski, L., Wissner, M., Hammer, S., Siefert, F., Seebach, H., Bernard, Y., Reif, W., Müller-Schloer, C., und André, E. (2013). Reference Architectures for Trustworthy Energy Management, Desktop Grid Computing Applications, and Ubiquitous Display Environments. Technischer Bericht 5, Universität Augsburg.

[Arnetz und Theorell, 1983] Arnetz, B. und Theorell, T. (1983). Psychological, Sociological and Health Behaviour Aspects of a Long Term Activation Programme for Institutionalized Elderly People. Social Science and Medicine, 17:449–456.

[Arning und Ziefle, 2007] Arning, K. und Ziefle, M. (2007). Understanding Age Differences in PDA Acceptance and Performance. Comput. Hum. Behav., 23(6):2904–2927.

[Asendorpf, 2015] Asendorpf, J. B. (2015). Persönlichkeitspsychologie für Bachelor. Springer-Verlag.

[Asendorpf und Neyer, 2012] Asendorpf, J. B. und Neyer, F. J. (2012). Psychologie der Persönlichkeit. Springer-Verlag.

[Azaria und Hong, 2016] Azaria, A. und Hong, J. (2016). Recommender Systems with Personality. In Proceedings of the 10th ACM Conference on Recommender Systems, RecSys '16, Seiten 207–210, New York, NY, USA. ACM.

[Bader et al., 2011a] Bader, R., Karitnig, A., Woerndl, W., und Leitner, G. (2011a). Explanations in Proactive Recommender Systems in Automotive Scenarios. Joint Proc. of the Workshop on Decision Making and Recommendation Acceptance Issues in Recommender Systems and the 2nd Workshop on User Models for Motivational Systems: The Affective and the Rational Routes to Persuasion., 740:11–18.

[Bader et al., 2011b] Bader, R., Siegmund, O., und Woerndl, W. (2011b). A Study on User Acceptance of Proactive In-vehicle Recommender Systems. In Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '11, Seiten 47–54, New York, NY, USA. ACM.

[Bader et al., 2010] Bader, R., Woerndl, W., und Prinz, V. (2010). Situation awareness for proactive in-car recommendations of points-of-interest (POI). In Workshop on Context-Aware Intelligent Assistance, Karlsruhe, Germany.

[Baltrunas et al., 2011] Baltrunas, L., Ludwig, B., Peer, S., und Ricci, F. (2011). Context-Aware Places of Interest Recommendations for Mobile Users. In Marcus, A., Herausgeber, Design, User Experience, and Usability. Theory, Methods, Tools and Practice, Band 6769 in Lecture Notes in Computer Science, Seiten 531–540. Springer Berlin Heidelberg.

[Barkhuus und Dey, 2003] Barkhuus, L. und Dey, A. (2003). Is Context-Aware Computing Taking Control away from the User? Three Levels of Interactivity Examined. In Dey, A. K., Schmidt, A., und McCarthy, J. F., Herausgeber, Proceedings of the 5th International Conference on Ubiquitous Computing (UbiComp 2003), Seiten 149–156. Springer, Berlin, Heidelberg.

[Bee et al., 2012] Bee, K., Hammer, S., Pratsch, C., und André, E. (2012). The Automatic Trust Management of Self-Adaptive Multi-Display Environments. In Trustworthy Ubiquitous Computing, Band 6 in Atlantis Ambient and Pervasive Intelligence, Seiten 3–20. Atlantis Press.

[Bee et al., 2010] Bee, N., Pollock, C., André, E., und Walker, M. (2010). Bossy or Wimpy: Expressing Social Dominance by Combining Gaze and Linguistic Behaviors. In Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., und Safonova, A., Herausgeber, Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings, Seiten 265–271. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Benders et al., 2006] Benders, R. M., Kok, R., Moll, H. C., Wiersma, G., und Noorman, K. J. (2006). New approaches for household energy conservation - in search of personal household energy budgets and energy reduction options. Energy Policy, 34(18):3612 – 3622.

[Bentley et al., 2013] Bentley, F., Tollmar, K., Stephenson, P., Levy, L., Jones, B., Robertson, S., Price, E., Catrambone, R., und Wilson, J. (2013). Health Mashups: Presenting Statistical Patterns Between Wellbeing Data and Context in Natural Language to Promote Behavior Change. ACM Trans. Comput.-Hum. Interact., 20(5):30:1–30:27.

[Berdichevsky und Neuenschwander, 1999] Berdichevsky, D. und Neuenschwander, E. (1999). Toward an Ethics of Persuasive Technology. Commun. ACM, 42(5):51–58.

[Bhuiyan et al., 2010] Bhuiyan, T., Xu, Y., und Jøsang, A. (2010). A Review of Trust in Online Social Networks to Explore New Research Agenda. In Arabnia, H. R., Clincy, V. A., Lu, J., Marsh, A., und Solo, A. M. G., Herausgeber, Proc. of the 2010 Int. Conf. on Internet Computing, ICOMP 2010, July 12-15, 2010, Seiten 123–128, Las Vegas, NV, USA. CSREA Press.

[Bilgic, 2005] Bilgic, M. (2005). Explaining Recommendations: Satisfaction vs. Promotion. In In Proceedings of Beyond Personalization 2005, the Workshop on the Next Stage of Recommender Systems Research(IUI2005, Seiten 13–18.

[Bogomolov, 2015] Bogomolov, S. (2015). Entwicklung eines Konzeptes für einen Systemen-Assistenten für ältere Nutzer unter Beachtung von Höflichkeitsstrategien in einem Empfehlungsszenario. Masterarbeit, betreut durch Hammer, S., Lugrin, B. und André, E., Universität Augsburg, Augsburg, Germany.

[Breese et al., 1998] Breese, J. S., Heckerman, D., und Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI'98, Seiten 43–52, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Briggs und Scheutz, 2016] Briggs, G. und Scheutz, M. (2016). The Pragmatic Social Robot: Toward Socially-Sensitive Utterance Generation in Human-Robot Interactions. In AAAI Fall Symposium Series: Artificial Intelligence for Human-Robot Interaction, Seiten 12–15, Arlington, VA, USA.

[Brown und Levinson, 1987] Brown, P. und Levinson, S. (1987). Politeness: Some Universals in Language Usage. Studies in Interactional Sociolinguistics. Cambridge University Press.

[Bühling et al., 2012] Bühling, R., Obaid, M., Hammer, S., und André, E. (2012). Mobile Augmented Reality and Adaptive Art: A Game-based Motivation for Energy Saving. In Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, MUM '12, Seiten 50:1–50:2, New York, NY, USA. ACM.

[Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit, 2016] Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit (2016). Themenbereich Energieeffizienz - Stromspartipps. `http://www.bmub.bund.de/themen/klima-energie/energieeffizienz/foerdermittel-beratung/stromspartipps/`.

[Burke, 2000] Burke, R. (2000). Knowledge-based Recommender Systems. Encyclopedia of Library and Information Science, 69(Supplement 32):180–200.

[Burke, 2002] Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, 12(4):331–370.

[Caballero et al., 2013] Caballero, F. F., Miret, M., Power, M., Chatterji, S., Tobiasz-Adamczyk, B., Koskinen, S., Leonardi, M., Olaya, B., Haro, J. M., und Ayuso-Mateos, J. L. (2013). Validation of an instrument to evaluate quality of life in the aging population: WHOQOL-AGE. Health and Quality of Life Outcomes, 11(1):177.

[Carenini und Moore, 2006] Carenini, G. und Moore, J. D. (2006). Generating and Evaluating Evaluative Arguments. Artif. Intell., 170(11):925–952.

[Carenini et al., 2003] Carenini, G., Smith, J., und Poole, D. (2003). Towards More Conversational and Collaborative Recommender Systems. In Proceedings of the 8th International Conference on Intelligent User Interfaces, IUI '03, Seiten 12–18, New York, NY, USA. ACM.

[Cassell und Bickmore, 2003] Cassell, J. und Bickmore, T. (2003). Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. User Modeling and User-Adapted Interaction, 13(1):89–132.

[Castelfranchi und Falcone, 2010] Castelfranchi, C. und Falcone, R. (2010). Trust Theory: A Socio-Cognitive and Computational Model. Wiley.

[Chaiken et al., 1996] Chaiken, S., Wood, W., und Eagly, A. H. (1996). Principles of persuasion. In Higgins, E. und Kruglanski, A., Herausgeber, Social psychology: Handbook of basic principles, Seiten 702–742. The Guilford Press: New York.

[Charness und Boot, 2009] Charness, N. und Boot, W. R. (2009). Aging and Information Technology Use: Potential and Barriers. Current Directions in Psychological Science, 18(5):253–258.

[Chen und Pu, 2012] Chen, L. und Pu, P. (2012). Critiquing-based recommenders: survey and emerging trends. User Modeling and User-Adapted Interaction, 22(1):125–150.

[Cheverst et al., 2005] Cheverst, K., Byun, H., Fitton, D., Sas, C., Kray, C., und Villar, N. (2005). Exploring Issues of User Model Transparency and Proactive Behaviour in an Office Environment Control System. User Modeling and User-Adapted Interaction, 15(3-4):235–273.

[Cialdini et al., 1981] Cialdini, R. B., Petty, R. E., und Cacioppo, J. T. (1981). Attitude and Attitude Change. Annual Review of Psychology, 32(1):357–404.

[co2 online, 2017] co2 online (2017). EnergiesparChecks. `https://www.co2online.de/service/energiesparchecks/`.

[Consolvo et al., 2008] Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., Smith, I., und Landay, J. A. (2008). Activity Sensing in the Wild: A Field Trial of Ubifit Garden. In Proc. of the SIGCHI Conf. on Human Factors in Comput. Systems, Seiten 1797–1806, New York, NY, USA. ACM.

[Consolvo und Towle, 2005] Consolvo, S. und Towle, J. (2005). Evaluating an Ambient Display for the Home. In CHI '05 Extended Abstracts on Human Factors in Computing Systems, CHI EA '05, Seiten 1304–1307, New York, NY, USA. ACM.

[Costa und MacCrae, 1992] Costa, P. T. und MacCrae, R. R. (1992). Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual. Psychological Assessment Resources.

[Cramer et al., 2008] Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., und Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-Adapted Interaction, 18(5):455.

[Davis, 1989] Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Q., 13(3):319–340.

[Davis und Venkatesh, 2004] Davis, F. D. und Venkatesh, V. (2004). Toward preprototype user acceptance testing of new information systems: implications for software project management. IEEE Transactions on Engineering Management, 51(1):31–46.

[Denko et al., 2011] Denko, M. K., Sun, T., und Woungang, I. (2011). Trust Management in Ubiquitous Computing: A Bayesian Approach. Computer Communications, 34(3):398–406.

[Deutsches Institut für Normung e.V., 2010] Deutsches Institut für Normung e.V. (2010). Ergonomie der Mensch-System-Interaktion - Teil 210: Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme (ISO 9241-210:2010). BeuthVerlag, Berlin, Germany.

[Dhamija et al., 2006] Dhamija, R., Tygar, J. D., und Hearst, M. (2006). Why Phishing Works. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06, Seiten 581–590, New York, NY, USA. ACM.

[Dillon und Morris, 1996] Dillon, A. und Morris, M. (1996). User Acceptance of Information Technology: Theories and Models. Annual Review of Information Science and Technology, 31:3–32.

[DiSalvo et al., 2010] DiSalvo, C., Sengers, P., und Brynjarsdóttir, H. (2010). Mapping the landscape of sustainable HCI. In Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, CHI '10, Seiten 1975–1984, New York, NY, USA. ACM.

[Dunn et al., 2009] Dunn, G., Wiersema, J., Ham, J., und Aroyo, L. (2009). Evaluating Interface Variants on Personality Acquisition for Recommender Systems. In Houben, G.-J., McCalla, G., Pianesi, F., und Zancanaro, M., Herausgeber, User Modeling, Adaptation, and Personalization: 17th International Conference, UMAP 2009, formerly UM and AH, Trento, Italy, June 22-26, 2009. Proceedings, Seiten 259–270. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Dzindolet et al., 2003] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., und Beck, H. P. (2003). The Role of Trust in Automation Reliance. Int. J. Hum.-Comput. Stud., 58(6):697–718.

[Ehrlich et al., 2011] Ehrlich, K., Kirk, S. E., Patterson, J., Rasmussen, J. C., Ross, S. I., und Gruen, D. M. (2011). Taking Advice from Intelligent Systems: The Double-edged Sword of Explanations. In Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI '11, Seiten 125–134, New York, NY, USA. ACM.

[Endrass et al., 2013] Endrass, B., André, E., Rehm, M., und Nakano, Y. (2013). Investigating culture-related aspects of behavior for virtual characters. Autonomous Agents and Multi-Agent Systems, 27(2):277–304.

[Erndl, 1998] Erndl, R. (1998). Höflichkeit im Deutschen - Konzeption zur Integration einer zentralen Gesprächskompetenz im Deutsch als Fremdsprache-Unterricht. Materialien Deutsch als Fremdsprache. 49. Fachverb. Deutsch als Fremdsprache, Regensburg.

[Eysenck und Eysenck, 1965] Eysenck, H. J. und Eysenck, S. G. B. (1965). The Eysenck Personality Inventory. British Journal of Educational Studies, 14(1).

[Felfernig et al., 2008] Felfernig, A., Gula, B., Leitner, G., Maier, M., Melcher, R., und Teppan, E. (2008). Persuasion in Knowledge-Based Recommendation. In Oinas-Kukkonen, H., Hasle, P., Harjumaa, M., Segerståhl, K., und Øhrstrøm, P., Herausgeber, Persuasive Technology, Band 5033 in LNCS, Seiten 71–82. Springer Berlin Heidelberg.

[Felfernig et al., 2013] Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F., und Reiterer, S. (2013). Toward the Next Generation of Recommender Systems: Applications and Research Challenges. In Tsihrintzis, A. G., Virvou, M., und Jain, C. L., Herausgeber, Multimedia Services in Intelligent Environments: Advances in Recommender Systems, Seiten 81–98. Springer International Publishing, Heidelberg.

[Flandorfer, 2012] Flandorfer, P. (2012). Population Ageing and Socially Assistive Robots for Elderly Persons: The Importance of Sociodemographic Factors for User Acceptance. International Journal of Population Research, 2012:13.

[Fogg, 2002] Fogg, B. (2002). Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Fogg, 2009] Fogg, B. (2009). A Behavior Model for Persuasive Design. In Proceedings of the 4th International Conference on Persuasive Technology, Persuasive '09, Seiten 40:1–40:7, New York, NY, USA. ACM.

[Ford et al., 2014] Ford, R., Sumavsk, O., Clarke, A., und Thorsnes, P. (2014). Personalized Energy Priorities: A User-Centric Application for Energy Advice. In Marcus, A., Herausgeber, Design, User Experience, and Usability. User Experience Design for Everyday Life Applications and Services, Band 8519 in LNCS, Seiten 542–553. Springer International Publishing.

[Fraser, 2001] Fraser, B. (2001). The Form and Function of Politeness in Conversation. In Brinker, K., Antos, G. Heinemann, W., und Sager, S., Herausgeber, Text- und Gesprächslinguistik, 2. Halbband, HSK, Seiten 1406–1425. de Gruyter, Berlin, Germany.

[Fraser und Nolen, 1981] Fraser, B. und Nolen, W. (1981). The Association of Deference with Linguistic FWorm. International Journal of the Sociology of Language, 27:93–110.

[Friedman, 1996] Friedman, B. (1996). Value-sensitive Design. interactions, 3(6):16–23.

[Friedman und Kahn, 2003] Friedman, B. und Kahn, Jr., P. H. (2003). Human Values, Ethics, and Design. In Jacko, J. A. und Sears, A., Herausgeber, The Human-computer Interaction Handbook, Seiten 1177–1201. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

[Friedrich und Zanker, 2011] Friedrich, G. und Zanker, M. (2011). A Taxonomy for Generating Explanations in Recommender Systems. AI Magazine, 32(3):90–98.

[Froehlich et al., 2009] Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., Harrison, B., und Landay, J. A. (2009). UbiGreen: Investigating a Mobile Tool for Tracking and Supporting Green Transportation Habits. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09, Seiten 1043–1052, New York, NY, USA. ACM.

[Furnham, 1990] Furnham, A. (1990). Language and Personality. In Giles, H. und Robinson, W. P., Herausgeber, Handbook of Language and Social Psychologie, Seiten 73–95. John Wiley & Sons, Oxford, UK.

[Galton, 1949] Galton, F. (1949). The Measurement of Character. In Wayne, D., Herausgeber, Readings in General Psychology, Seiten 435–444. Prentice-Hall, Inc, New York, NY, US.

[Gamberini et al., 2012] Gamberini, L., Spagnolli, A., Corradi, N., Jacucci, G., Tusa, G., Mikkola, T., Zamboni, L., und Hoggan, E. (2012). Tailoring Feedback to Users' Actions in a Persuasive Game for Household Electricity Conservation. In Bang, M. und Ragnemalm, E., Herausgeber, Persuasive Technology. Design for Health and Safety, Band 7284 in Lecture Notes in Computer Science, Seiten 100–111. Springer Berlin Heidelberg.

[Gardner und Stern, 2008] Gardner, G. T. und Stern, P. C. (2008). The Short List: The Most Effective Actions U.S. Households Can Take to Curb Climate Change. Environment: Science and Policy for Sustainable Development, 50(5):12–25.

[Ge et al., 2010] Ge, M., Delgado-Battenfeld, C., und Jannach, D. (2010). Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10, Seiten 257–260, New York, NY, USA. ACM.

[Gebhard et al., 2012] Gebhard, P., Mehlmann, G., und Kipp, M. (2012). Visual SceneMaker - A Tool for Authoring Interactive Virtual Characters. Multimodal User Interfaces, 6(1-2):3–11.

[Gedikli et al., 2014] Gedikli, F., Jannach, D., und Ge, M. (2014). How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. Int. J. Hum.-Comput. Stud., 72(4):367–382.

[Gefen et al., 2003] Gefen, D., Karahanna, E., und Straub, D. W. (2003). Trust and TAM in Online Shopping: An Integrated Model. MIS Q., 27(1):51–90.

[Gkika und Lekakos, 2014] Gkika, S. und Lekakos, G. (2014). The persuasive role of Explanations in Recommender Systems. In 2nd Int. Workshop on Behavior Change Support Systems (BCSS 2014), Band 1153, Seiten 59–68.

[Glass et al., 2008] Glass, A., McGuinness, D. L., und Wolverton, M. (2008). Toward Establishing Trust in Adaptive Agents. In Proc. of the 13th Int. Conf. on Intelligent User Interfaces (IUI '08), Seiten 227–236. ACM.

[Golbeck, 2005] Golbeck, J. A. (2005). Computing and Applying Trust in Web-based Social Networks. Dissertation, University of Maryland at College Park, College Park, MD, USA. AAI3178583.

[Goldberg et al., 1992] Goldberg, D., Nichols, D., Oki, B. M., und Terry, D. (1992). Using Collaborative Filtering to Weave an Information Tapestry. Commun. ACM, 35(12):61–70.

[Goldberg et al., 2001] Goldberg, K., Roeder, T., Gupta, D., und Perkins, C. (2001). Eigentaste: A Constant Time Collaborative Filtering Algorithm. Information Retrieval, 4(2):133–151.

[Gosling et al., 2003] Gosling, S. D., Rentfrow, P. J., und Jr., W. B. S. (2003). A very brief measure of the Big-Five personality domains. Journal of Research in Personality, 37(6):504 – 528.

[Gregor und Benbasat, 1999] Gregor, S. und Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. MIS Q., 23(4):497–530.

[Grice, 1975] Grice, H. P. (1975). Logic and Conversation. In Cole, P. und Morgan, J. L., Herausgeber, Syntax and Semantics: Vol. 3: Speech Acts, Seiten 41–58. Academic Press, San Diego, CA.

[Gustafsson et al., 2010] Gustafsson, A., Katzeff, C., und Bang, M. (2010). Evaluation of a Pervasive Game for Domestic Energy Engagement Among Teenagers. Comput. Entertain., 7(4):54:1–54:19.

[Hamari et al., 2014] Hamari, J., Koivisto, J., und Sarsa, H. (2014). Does Gamification Work? – A Literature Review of Empirical Studies on Gamification. In 2014 47th Hawaii International Conference on System Sciences, Seiten 3025–3034.

[Hammer et al., 2013] Hammer, S., Kiefhaber, R., Redlin, M., André, E., und Ungerer, T. (2013). A User-Centric Study Of Reputation Metrics in Online Communities. In Proc. of the 3rd Workshop on Trust, Reputation and User Modeling (TRUM'13).

[Hammer et al., 2010] Hammer, S., Kim, J., und André, E. (2010). MED-StyleR: METABO Diabetes-Lifestyle Recommender. In Proc. of the 4th ACM Conf. on Recommender Systems, RecSys '10, Seiten 285–288, New York, NY, USA. ACM.

[Hammer et al., 2017] Hammer, S., Kirchner, K., André, E., und Lugrin, B. (2017). Touch or Talk?: Comparing Social Robots and Tablet PCs for an Elderly Assistant Recommender System. In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI '17, Seiten 129–130, New York, NY, USA. ACM.

[Hammer et al., 2016a] Hammer, S., Lugrin, B., Bogomolov, S., Janowski, K., und André, E. (2016a). Investigating Politeness Strategies and Their Persuasiveness for a Robotic Elderly Assistant. In Meschtscherjakov, A., De Ruyter, B., Fuchsberger, V., Murer, M., und Tscheligi, M., Herausgeber, Proc. of the 11th Int. Conf. on Persuasive Technology (PERSUASIVE 2016), Seiten 315–326, Cham. Springer International Publishing.

[Hammer et al., 2015a] Hammer, S., Segmüller, F., Lugrin, B., und André, E. (2015a). Promoting Energy-Efficient Behavior by Recommendations based on Energy Cultures. In Adjunct Conference Proceedings of INTERACT 2015, INTERACT Workshop on Fostering Smart Energy Applications.

[Hammer et al., 2015b] Hammer, S., Seiderer, A., André, E., Rist, T., Kastrinaki, S., Hondrou, C., Raouzaiou, A., Karpouzis, K., und Kollias, S. (2015b). Design of a Lifestyle Recommender System for the Elderly: Requirement Gatherings in Germany and Greece. In Proc. of the 8th ACM Int. Conf. on Pervasive Technologies Related to Assistive Environments, PETRA '15, Seiten 80:1–80:8, New York, NY, USA. ACM.

[Hammer et al., 2014] Hammer, S., Wißner, M., und André, E. (2014). Trust-Based Decision-Making for Energy-Aware Device Management. In Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., und Houben, G.-J., Herausgeber, User Modeling, Adaptation, and Personalization, Seiten 326–337, Cham. Springer International Publishing.

[Hammer et al., 2015c] Hammer, S., Wißner, M., und André, E. (2015c). Trust-based decision-making for smart and adaptive environments. User Modeling and User-Adapted Interaction, 25(3):267–293.

[Hammer et al., 2016b] Hammer, S., Wißner, M., und André, E. (2016b). A User Trust Model for Automatic Decision-Making in Ubiquitous and Self-Adaptive Environments. In Reif, W., Anders, G., Seebach, H., Steghöfer, J.-P., André, E., Hähner, J., Müller-Schloer, C., und Ungerer, T., Herausgeber, Trustworthy Open Self-Organising Systems, Seiten 55–87. Springer International Publishing, Cham.

[Hampton-Sosa und Koufaris, 2005] Hampton-Sosa, W. und Koufaris, M. (2005). The Effect of Web Site Perceptions on Initial Trust in the Owner Company. Int. J. Electron. Commerce, 10(1):55–81.

[Han und Shavitt, 1994] Han, S.-p. und Shavitt, S. (1994). Persuasion and Culture: Advertising Appeals in Individualistic and Collectivistic Societies. Journal of Experimental Social Psychology, 30(4):326 – 350.

[Hazas et al., 2011] Hazas, M., Friday, A., und Scott, J. (2011). Look Back before Leaping Forward: Four Decades of Domestic Energy Inquiry. Pervasive Computing, IEEE, 10(1):13–19.

[He et al., 2010] He, H. A., Greenberg, S., und Huang, E. M. (2010). One Size Does Not Fit All: Applying the Transtheoretical Model to Energy Feedback Technology Design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10, Seiten 927–936, New York, NY, USA. ACM.

[Heerink, 2010] Heerink, M. (2010). Assessing acceptance of assistive social robots by aging adults. Dissertation, University of Amsterdam.

[Heerink, 2011] Heerink, M. (2011). Exploring the Influence of Age, Gender, Education and Computer Experience on Robot Acceptance by Older Adults. In Proceedings of the 6th International Conference on Human-robot Interaction, HRI '11, Seiten 147–148, New York, NY, USA. ACM.

[Heijden, 2004] Heijden, H. (2004). User Acceptance of Hedonic Information Systems. MIS Q., 28(4):695–704.

[Heizsparer.de, 2017] Heizsparer.de (2017). Heizungssysteme - Verschiedene Heizungssysteme ausführlich vorgestellt. http://www.heizsparer.de/heizung/heizungssysteme.

[Herlocker et al., 2000] Herlocker, J. L., Konstan, J. A., und Riedl, J. (2000). Explaining Collaborative Filtering Recommendations. In Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, CSCW '00, Seiten 241–250, New York, NY, USA. ACM.

[Herlocker et al., 2004] Herlocker, J. L., Konstan, J. A., Terveen, L. G., und Riedl, J. T. (2004). Evaluating Collaborative Filtering Recommender Systems. ACM Trans. Inf. Syst., 22(1):5–53.

[Hoens et al., 2013] Hoens, T. R., Blanton, M., Steele, A., und Chawla, N. V. (2013). Reliable Medical Recommendation Systems with Patient Privacy. ACM Trans. Intell. Syst. Technol., 4(4):67:1–67:31.

[Hofstede, 2001] Hofstede, G. (2001). Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations. Second Edition. Sage Publications, Thousand Oaks, CA, USA.

[Hofstede, 2017] Hofstede, G. (2017). Geert Hofstede - Country comparison. https://geert-hofstede.com/countries.html.

[Hofstede et al., 2010] Hofstede, G., Hofstede, G. J., und Minkov, M. (2010). Cultures and Organizations: Software of the Mind: Intercultural Cooperation and its Importance for Survival. Revised and Expanded 3rd Edition. McGraw-Hill, New York, NY, USA.

[Höök, 1997] Höök, K. (1997). Evaluating the Utility and Usability of an Adaptive Hypermedia System. In Proceedings of the 2Nd International Conference on Intelligent User Interfaces, IUI '97, Seiten 179–186, New York, NY, USA. ACM.

[Höök, 2000] Höök, K. (2000). Steps to take before intelligent user interfaces become real. Interacting with Computers, 12(4):409–426.

[Horvitz, 1999] Horvitz, E. (1999). Principles of Mixed-initiative User Interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '99, Seiten 159–166, New York, NY, USA. ACM.

[Horvitz et al., 2003] Horvitz, E., Kadie, C., Paek, T., und Hovel, D. (2003). Models of Attention in Computing and Communication: From Principles to Applications. Commun. ACM, 46(3):52–59.

[Hu und Pu, 2009] Hu, R. und Pu, P. (2009). Acceptance Issues of Personality-based Recommender Systems. In Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, Seiten 221–224, New York, NY, USA. ACM.

[Hu und Pu, 2011] Hu, R. und Pu, P. (2011). Enhancing Collaborative Filtering Systems with Personality Information. In Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, Seiten 197–204, New York, NY, USA. ACM.

[Hutterer, 2013] Hutterer, R. (2013). Das Paradigma der Humanistischen Psychologie: Entwicklung, Ideengeschichte und Produktivität. Springer-Verlag.

[Ikemoto et al., 2018] Ikemoto, Y., Asawavetvutt, V., Kuwabara, K., und Huang, H.-H. (2018). Conversation Strategy of a Chatbot for Interactive Recommendations. In Nguyen, N. T., Hoang, D. H., Hong, T.-P., Pham, H., und Trawiński, B., Herausgeber, Intelligent Information and Database Systems, Seiten 117–126, Cham. Springer International Publishing.

[Ivanov et al., 2013] Ivanov, I., Vajda, P., Korshunov, P., und Ebrahimi, T. (2013). Comparative Study of Trust Modeling for Automatic Landmark Tagging. IEEE Transactions on Information Forensics and Security, 8(6):911–923.

[Jameson, 2003] Jameson, A. (2003). Adaptive Interfaces and Agents. In Jacko, J. A. und Sears, A., Herausgeber, The Human-computer Interaction Handbook, Seiten 305–330. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

[Jannach et al., 2015] Jannach, D., Lerche, L., und Jugovac, M. (2015). Item Familiarity as a Possible Confounding Factor in User-Centric Recommender Systems Evaluation. i-com, 14(1):29–39.

[Jannach et al., 2016] Jannach, D., Resnick, P., Tuzhilin, A., und Zanker, M. (2016). Recommender Systems - Beyond Matrix Completion. Commun. ACM, 59(11):94–102.

[Jannach et al., 2010] Jannach, D., Zanker, M., Felfernig, A., und Friedrich, G. (2010). Recommender Systems: An Introduction. Cambridge University Press, New York, NY, USA, 1st. Auflage.

[Johnson und Eagly, 1989] Johnson, B. T. und Eagly, A. H. (1989). Effects of involvement on persuasion: A meta-analysis. Psychological bulletin, 106(2):290–314.

[Johnson et al., 2005] Johnson, W. L., Mayer, R. E., André, E., und Rehm, M. (2005). Cross-Cultural Evaluation of Politeness in Tactics for Pedagogical Agents. In Proc. of the 2005 Conference on Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology, Seiten 298–305, Amsterdam, Netherlands. IOS Press.

[Jones und Pu, 2008] Jones, N. und Pu, P. (2008). User Acceptance Issues in Music Recommender Systems. Technischer bericht, EPFL.

[Jøsang und Presti, 2004] Jøsang, A. und Presti, S. L. (2004). Analysing the Relationship between Risk and Trust. In Jensen, C., Poslad, S., und Dimitrakos, T., Herausgeber, Trust Management: Second International Conference, iTrust 2004, Oxford, UK, March 29 - April 1, 2004. Proceedings, Seiten 135–145. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Jung und Baynes, 1923] Jung, C. G. und Baynes, H. G. (1923). Psychological Types. Journal of Philosophy, 20(23):636–640.

[Kaasinen, 2005] Kaasinen, E. (2005). User acceptance of mobile services – value, ease of use, trust and ease of adoption. Dissertation, VTT Technical Research Centre of Finland.

[Kaminskas und Bridge, 2016] Kaminskas, M. und Bridge, D. (2016). Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. ACM Trans. Interact. Intell. Syst., 7(1):2:1–2:42.

[Kaneda et al., 2011] Kaneda, T., Lee, M., und Pollard, K. (2011). SCL/PRB index of well-being in older populations. Washington DC: Population Reference Bureau.

[Karumur et al., 2016] Karumur, R. P., Nguyen, T. T., und Konstan, J. A. (2016). Exploring the Value of Personality in Predicting Rating Behaviors: A Study of Category Preferences on MovieLens. In Proceedings of the 10th ACM Conference on Recommender Systems, RecSys '16, Seiten 139–142, New York, NY, USA. ACM.

[Kay, 2006] Kay, J. (2006). Scrutable Adaptation: Because We Can and Must. In Wade, V. P., Ashman, H., und Smyth, B., Herausgeber, Adaptive Hypermedia and Adaptive Web-Based Systems: 4th International Conference, AH 2006, Dublin, Ireland, June 21-23, 2006. Proceedings, Seiten 11–19. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Khaled et al., 2006] Khaled, R., Biddle, R., Noble, J., Barr, P., und Fischer, R. (2006). Persuasive Interaction for Collectivist Cultures. In Proceedings of the 7th Australasian User Interface Conference - Volume 50, AUIC '06, Seiten 73–80, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

[Kiefhaber et al., 2011] Kiefhaber, R., Hammer, S., Savs, B., Schmitt, J., Roth, M., Kluge, F., Andre, E., und Ungerer, T. (2011). The Neighbor-Trust Metric to Measure Reputation in Organic Computing Systems. In 2011 Fifth IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshops, Seiten 41–46.

[Kientz et al., 2010] Kientz, J. A., Choe, E. K., Birch, B., Maharaj, R., Fonville, A., Glasson, C., und Mundt, J. (2010). Heuristic Evaluation of Persuasive Health Technologies. In Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10, Seiten 555–564, New York, NY, USA. ACM.

[Knijnenburg et al., 2013] Knijnenburg, B. P., Kobsa, A., und Jin, H. (2013). Dimensionality of information disclosure behavior. Int. J. Hum.-Comput. Stud., 71(12):1144–1162.

[Knijnenburg et al., 2012] Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., und Newell, C. (2012). Explaining the User Experience of Recommender Systems. User Modeling and User-Adapted Interaction, 22(4-5):441–504.

[Komiak und Benbasat, 2006] Komiak, S. Y. X. und Benbasat, I. (2006). The Effects of Personalizaion and Familiarity on Trust and Adoption of Recommendation Agents. MIS Q., 30(4):941–960.

[Konstan und Riedl, 2012] Konstan, J. A. und Riedl, J. (2012). Recommender Systems: From Algorithms to User Experience. User Modeling and User-Adapted Interaction, 22(1):101–123.

[Koren et al., 2009] Koren, Y., Bell, R., und Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. Computer, 42(8):30–37.

[Koufaris und Hampton-Sosa, 2002] Koufaris, M. und Hampton-Sosa, W. (2002). Customer trust online: examining the role of the experience with the Web-site. Department of Statistics and Computer Information Systems Working Paper Series, Zicklin School of Business, Baruch College, New York.

[Krenn et al., 2014] Krenn, B., Endrass, B., Kistler, F., und André, E. (2014). Effects of Language Variety on Personality Perception in Embodied Conversational Agents. In Kurosu, M., Herausgeber, Human-Computer Interaction. Advanced Interaction Modalities and Techniques: 16th International Conference, HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part II, Seiten 429–439. Springer International Publishing, Cham.

[Krulwich, 1997] Krulwich, B. (1997). LIFESTYLE FINDER: Intelligent User Profiling Using Large-Scale Demographic Data. AI Magazine, 18(2):37–45.

[Kurdyukova et al., 2012] Kurdyukova, E., Hammer, S., und André, E. (2012). Personalization of Content on Public Displays Driven by the Recognition of Group Context. In Paternò, F., de Ruyter, B., Markopoulos, P., Santoro, C., van Loenen, E., und Luyten, K., Herausgeber, Ambient Intelligence, Band 7683 in Lecture Notes in Computer Science, Seiten 272–287. Springer Berlin Heidelberg.

[Kurdyukova et al., 2013] Kurdyukova, E., Wissner, M., Hammer, S., und André, E. (2013). Trust-based decision-making for the adaptation of public displays in changing social contexts. In 2013 Eleventh Annual Conference on Privacy, Security and Trust, Seiten 317–324.

[Labov, 1984] Labov, W. (1984). Field methods of the Project on Linguistic Change and Variation. In Baugh, J. und Sherzer, J., Herausgeber, Language in use: Readings in sociolinguistics, Seiten 28–53. Prentice Hall, Englewood Cliffs, NJ, USA.

[Lakoff, 1973] Lakoff, R. (1973). The logic of politeness: or minding your p's and q's. Papers from the Ninth Regional Meeting of the Chicago Linguistic Society, Seiten 292–305.

[Lane et al., 2014] Lane, N. D., Lin, M., Mohammod, M., Yang, X., Lu, H., Cardone, G., Ali, S., Doryab, A., Berke, E., Campbell, A. T., und Choudhury, T. (2014). BeWell: Sensing Sleep, Physical Activities and Social Interactions to Promote Wellbeing. Mobile Networks and Applications, 19(3):345–359.

[Lang und Lüdtke, 2005] Lang, F. R. und Lüdtke, O. (2005). Der Big Five-Ansatz der Persönlichkeitsforschung: Instrumente und Vorgehen. In Persönlichkeit: eine vergessene Größe der emprischen Sozialforschung, Seiten 29–39. VS Verlag für Sozialwissenschaften.

[Langkilde und Knight, 1998] Langkilde, I. und Knight, K. (1998). Generation That Exploits Corpus-based Statistical Knowledge. In Proceedings of the 17th International Conference on Computational Linguistics - Volume 1, COLING '98, Seiten 704–710, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Langkilde-Geary, 2002] Langkilde-Geary, I. (2002). An empirical verification of coverage and correctness for a general-purpose sentence generator. In Proceedings of the 1st International Natural Language Generation Conference, Seiten 17–24.

[Lawson und Williams, 2012] Lawson, R. und Williams, J. (2012). Understanding energy cultures. In conference of the Australian and New Zealand Marketing Academy, Seiten 3–5.

[Lee und See, 2004] Lee, J. D. und See, K. A. (2004). Trust in automation: designing for appropriate reliance. Human Factors, 46(1):50–80.

[Lee und Ashton, 2004] Lee, K. und Ashton, M. C. (2004). Psychometric Properties of the HEXACO Personality Inventory. Multivariate Behavioral Research, 39(2):329–358.

[Leech, 1983] Leech, G. N. (1983). Principles of Pragmatics. Longman, London.

[Leichtenstern et al., 2010] Leichtenstern, K., André, E., und Kurdyukova, E. (2010). Managing User Trust for Self-adaptive Ubiquitous Computing Systems. In Proceedings of the 8th International Conference on Advances in Mobile Computing and Multimedia, MoMM '10, Paris, France, Seiten 409–414. ACM.

[Leichtenstern et al., 2011] Leichtenstern, K., Bee, N., André, E., Berkmüller, U., und Wagner, J. (2011). Physiological Measurement of Trust-Related Behavior in Trust-Neutral and Trust-Critical Situations. In Wakeman, I., Gudes, E., Jensen, C. D., und Crampton, J., Herausgeber, Trust Management V: 5th IFIP WG 11.11 International Conference, IFIPTM 2011, Copenhagen, Denmark, June 29 – July 1, 2011. Proceedings, Seiten 165–172. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Lin und McLeod, 2002] Lin, C.-H. und McLeod, D. (2002). Exploiting and learning human temperaments for customized information recommendation. In Proceedings of the 6th IASTED International Conference on Internet and Multimedia Systems and Applications. ACTA. iv+430.

[Lin et al., 2011] Lin, Y., Jessurun, J., de Vries, B., und Timmermans, H. (2011). Motivate: Context Aware Mobile Application for Activity Recommendation. In Proc. of the 2nd Int. Conf. on Ambient Intelligence, Seiten 210–214, Berlin, Heidelberg. Springer-Verlag.

[Linden et al., 2003] Linden, G., Smith, B., und York, J. (2003). Amazon.Com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing, 7(1):76–80.

[Littell und Girvin, 2002] Littell, J. H. und Girvin, H. (2002). Stages of Change. Behavior Modification, 26(2):223–273. PMID: 11961914.

[Lumsden, 2009] Lumsden, J. (2009). Triggering Trust: To what Extent does the Question Influence the Answer when Evaluating the Perceived Importance of Trust Triggers? In Proc. of the 2009 British Computer Society Conf. on Human-Computer Interaction (BCS HCI '09), Seiten 214–223. British Computer Society.

[Maes, 1994] Maes, P. (1994). Agents That Reduce Work and Information Overload. Commun. ACM, 37(7):30–40.

[Mahmood und Ricci, 2009] Mahmood, T. und Ricci, F. (2009). Improving Recommender Systems with Adaptive Conversational Strategies. In Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, HT '09, Seiten 73–82, New York, NY, USA. ACM.

[Mahncke et al., 2006] Mahncke, H. W., Bronstone, A., und Merzenich, M. M. (2006). Brain plasticity and functional losses in the aged: scientific bases for a novel intervention. In Møller, A. R., Herausgeber, Reprogramming of the Brain, Band 157 in Progress in Brain Research, Seiten 81 – 109. Elsevier.

[Mairesse, 2008] Mairesse, F. (2008). Learning to Adapt in Dialogue Systems: Data-driven Models for Personality Recognition and Generation. Dissertation, University of Sheffield, Department of Computer Science.

[Mairesse und Walker, 2007] Mairesse, F. und Walker, M. (2007). PERSONAGE: Personality Generation for Dialogue. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), Prague.

[Mairesse und Walker, 2010] Mairesse, F. und Walker, M. A. (2010). Towards Personality-based User Adaptation: Psychologically Informed Stylistic Language Generation. User Modeling and User-Adapted Interaction, 20(3):227–278.

[Mairesse et al., 2007] Mairesse, F., Walker, M. A., Mehl, M. R., und Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. J. Artif. Int. Res., 30(1):457–500.

[Mankoff et al., 2003] Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., und Ames, M. (2003). Heuristic Evaluation of Ambient Displays. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '03, Seiten 169–176, New York, NY, USA. ACM.

[Marangunić und Granić, 2015] Marangunić, N. und Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. Universal Access in the Information Society, 14(1):81–95.

[Marcu, 1996] Marcu, D. (1996). The Conceptual and Linguistic Facets of Persuasive Arguments. In ECAI Workshop - Gaps and Bridges: New Directions in Planning and Natural Language Generation, Seiten 43–46.

[Marsh, 1992] Marsh, S. (1992). Trust in Distributed Artificial Intelligence. In Castelfranchi, C. und Werner, E., Herausgeber, Artificial Social Systems, 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW '92, S. Martino al Cimino, Italy, July 29-31, 1992, Selected Papers, Band 830 in Lecture Notes in Computer Science, Seiten 94–112, Berlin, Heidelberg. Springer.

[Massa und Avesani, 2007a] Massa, P. und Avesani, P. (2007a). Trust-aware Recommender Systems. In Proceedings of the 2007 ACM Conference on Recommender Systems, RecSys '07, Seiten 17–24, New York, NY, USA. ACM.

[Massa und Avesani, 2007b] Massa, P. und Avesani, P. (2007b). Trust Metrics on Controversial Users: Balancing Between Tyranny of the Majority. Int. J. Semantic Web Inf. Syst., 3(1):39–64.

[Masthoff, 2011] Masthoff, J. (2011). Group Recommender Systems: Combining Individual Models. In Ricci, F., Rokach, L., Shapira, B., und Kantor, P. B., Herausgeber, Recommender Systems Handbook, Seiten 677–702. Springer US, Boston, MA.

[Mayer et al., 1995] Mayer, R. C., Davis, J. H., und Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. ACADEMY OF MANAGEMENT REVIEW, 20(3):709–734.

[McCrae, 1996] McCrae, R. R. (1996). Social consequences of experiential openness. Psychological bulletin, 120(3):323–337.

[McGinty und Reilly, 2011] McGinty, L. und Reilly, J. (2011). On the Evolution of Critiquing Recommenders. In Ricci, F., Rokach, L., Shapira, B., und Kantor, P. B., Herausgeber, Recommender Systems Handbook, Seiten 419–453. Springer US, Boston, MA.

[McLaughlin und Herlocker, 2004] McLaughlin, M. R. und Herlocker, J. L. (2004). A Collaborative Filtering Algorithm and Evaluation Metric That Accurately Model the User Experience. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, Seiten 329–336, New York, NY, USA. ACM.

[McNee, 2006] McNee, S. M. (2006). Meeting User Information Needs in Recommender Systems. Dissertation, University of Minnesota, Minneapolis, MN, USA. AAI3230139.

[McNee et al., 2006] McNee, S. M., Riedl, J., und Konstan, J. A. (2006). Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In CHI '06 Extended Abstracts on Human Factors in Computing Systems, CHI EA '06, Seiten 1097–1101, New York, NY, USA. ACM.

[McSweeney, 2002] McSweeney, B. (2002). Hofstede's Model of National Cultural Differences and their Consequences: A Triumph of Faith - a Failure of Analysis. Human Relations, 55(1):89–118.

[Mehlmann et al., 2015] Mehlmann, G., Janowski, K., und André, E. (2015). Modeling Grounding for Interactive Social Companions. KI - Künstliche Intelligenz, Seiten 1–8.

[Melguizo et al., 2007] Melguizo, M. C. P., Bogers, T., Deshpande, A., Boves, L., und van den Bosch, A. (2007). What a Proactive Recommendation System Needs - Relevance, Non-Intrusiveness, and a New Long-Term Memory. In ICEIS 2007 - Proceedings of the Ninth International Conference on Enterprise Information Systems, Seiten 86–91.

[Mertens, 2016] Mertens, J. (2016). Generierung kontextbewusster Empfehlungen und deren Präsentation innerhalb eines Dialogs auf Basis von Persönlichkeitsmerkmalen. Masterarbeit, betreut durch Hammer, S. und André, E., Universität Augsburg, Augsburg, Germany.

[Michaelson et al., 2009] Michaelson, J., Abdallah, S., Steuer, N., Thompson, S., Marks, N., Aked, J., Cordon, C., und Potts, R. (2009). National Accounts of Well-being: Bringing Real Wealth onto the Balance Sheet. `https://www.unicef.org/lac/National_Accounts_of_Well-being(1).pdf`.

[Middleton et al., 2009] Middleton, S. E., Roure, D. D., und Shadbolt, N. R. (2009). Ontology-Based Recommender Systems. In Staab, S. und Studer, R., Herausgeber, Handbook on Ontologies, Seiten 779–796. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Mitchell et al., 1994] Mitchell, T. M., Caruana, R., Freitag, D., McDermott, J., und Zabowski, D. (1994). Experience with a Learning Personal Assistant. Commun. ACM, 37(7):80–91.

[Motti et al., 2013] Motti, L. G., Vigouroux, N., und Gorce, P. (2013). Interaction Techniques for Older Adults Using Touchscreen Devices: A Literature Review. In Proc. of 25ème Conférence Francophone sur L'Interaction Homme-Machine, Seiten 125:125–125:134, New York, NY, USA. ACM.

[Nasoz et al., 2010] Nasoz, F., Lisetti, C. L., und Vasilakos, A. V. (2010). Affectively Intelligent and Adaptive Car Interfaces. Inf. Sci., 180(20):3817–3836.

[Nass und Lee, 2001] Nass, C. und Lee, K. M. (2001). Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology: Applied, 7:171–181.

[Nguyen et al., 2007] Nguyen, H., Masthoff, J., und Edwards, P. (2007). Modelling a Receiver's Position to Persuasive Arguments. In Proceedings of the 2nd International Conference on Persuasive Technology, PERSUASIVE'07, Seiten 271–282, Berlin, Heidelberg. Springer-Verlag.

[Nilashi et al., 2016] Nilashi, M., Jannach, D., Ibrahim, O. b., Esfahani, M. D., und Ahmadi, H. (2016). Recommendation Quality, Transparency, and Website Quality for Trust-building in Recommendation Agents. Electron. Commer. Rec. Appl., 19(C):70–84.

[Nomura und Saeki, 2010] Nomura, T. und Saeki, K. (2010). Effects of Polite Behaviors Expressed by Robots: A Psychological Experiment in Japan. Int. J. Synth. Emot., 1(2):38–52.

[Nomura und Takeuchi, 2011] Nomura, T. und Takeuchi, S. (2011). The Elderly and Robots: From Experiments Based on Comparison with Younger People. In Proceedings of the 12th AAAI Conference on Human-Robot Interaction in Elder Care, AAAIWS'11-12, Seiten 25–31. AAAI Press.

[Nunes, 2008] Nunes, M. A. S. N. (2008). Recommender systems based on personality traits. Dissertation, Université Montpellier II-Sciences et Techniques du Languedoc.

[Nunes, 2010] Nunes, M. A. S. N. (2010). Towards to Psychological-based Recommenders Systems: A survey on Recommender Systems. Scientia Plena, 6(8).

[O'Connor und Seymour, 2011] O'Connor, J. und Seymour, J. (2011). Introducing NLP: Psychological skills for understanding and influencing people. Conari Press.

[O'Donovan und Smyth, 2005] O'Donovan, J. und Smyth, B. (2005). Trust in Recommender Systems. In Proceedings of the 10th International Conference on Intelligent User Interfaces, Seiten 167–174, New York, NY, USA. ACM.

[Oinas-Kukkonen und Harjumaa, 2008] Oinas-Kukkonen, H. und Harjumaa, M. (2008). A Systematic Framework for Designing and Evaluating Persuasive Systems. In Oinas-Kukkonen, H., Hasle, P., Harjumaa, M., Segerståhl, K., und Øhrstrøm, P., Herausgeber, Persuasive Technology: Third International Conference, PERSUASIVE 2008, Oulu,

Finland, June 4-6, 2008. Proceedings, Seiten 164–176. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Oinas-Kukkonen und Harjumaa, 2009] Oinas-Kukkonen, H. und Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. Communications of the Association for Information Systems, 24(1):28.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., und Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. In Proceedings of the 7th International World Wide Web Conference, Seiten 161–172, Brisbane, Australia.

[Pavlou, 2003] Pavlou, P. A. (2003). Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model. Int. J. Electron. Commerce, 7(3):101–134.

[Pazzani, 1999] Pazzani, M. J. (1999). A Framework for Collaborative, Content-Based and Demographic Filtering. Artificial Intelligence Review, 13(5-6):393–408.

[Pennebaker und King, 1999] Pennebaker, J. W. und King, L. A. (1999). Linguistic styles: language use as an individual difference. Journal of personality and social psychology, 77(6):1296.

[Pervin et al., 2005] Pervin, L. A., Cervone, D., und John, O. P. (2005). Persönlichkeitstheorien. Ernst Reinhardt Verlag München Basel.

[Power et al., 1999] Power, M., Bullinger, M., und Harper, A. (1999). The World Health Organization WHOQOL-100: Tests of the universality of quality of life in 15 different cultural groups worldwide. Health psychology, 18(5):495.

[Prochaska, 2013] Prochaska, J. O. (2013). Transtheoretical Model of Behavior Change. In Gellman, M. D. und Turner, J. R., Herausgeber, Encyclopedia of Behavioral Medicine, Seiten 1997–2000. Springer New York, New York, NY.

[Prochaska und Velicer, 1997] Prochaska, J. O. und Velicer, W. F. (1997). The Transtheoretical Model of Health Behavior Change. American Journal of Health Promotion, 12(1):38–48. PMID: 10170434.

[Pu und Chen, 2006] Pu, P. und Chen, L. (2006). Trust Building with Explanation Interfaces. In Proc. of the 11th Int. Conf. on Intelligent User Interfaces, IUI '06, Seiten 93–100, New York, NY, USA. ACM.

[Pu et al., 2011] Pu, P., Chen, L., und Hu, R. (2011). A User-centric Evaluation Framework for Recommender Systems. In Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, Seiten 157–164, New York, NY, USA. ACM.

[Rafailidis und Crestani, 2017] Rafailidis, D. und Crestani, F. (2017). Learning to Rank with Trust and Distrust in Recommender Systems. In Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys '17, Seiten 5–13, New York, NY, USA. ACM.

[Reed und Long, 1997] Reed, C. und Long, D. (1997). Content Ordering in the Generation of Persuasive Discourse. In Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence - Volume 2, IJCAI'97, Seiten 1022–1027, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Reeves und Nass, 1998] Reeves, B. und Nass, C. (1998). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, New York, NY, USA.

[Resnick et al., 1994] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., und Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, CSCW '94, Seiten 175–186, New York, NY, USA. ACM.

[Resnick und Varian, 1997] Resnick, P. und Varian, H. R. (1997). Recommender Systems. Communications of the ACM, 40(3):56–58.

[RheinEnergie AG, 2017] RheinEnergie AG (2017). Stromverbrauchs-Check. `http://www.rheinenergie.com/de/privatkundenportal/energie___sicherheits_ _und_foerdermittelberatung/stromcheck/index.php`.

[Rhodes, 2000] Rhodes, B. J. (2000). Just-in-time Information Retrieval. Dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA. AAI0802535.

[Rich, 1979] Rich, E. (1979). User Modeling via Stereotypes. Cognitive Science, 3(4):329–354.

[Ring et al., 2015] Ring, L., Shi, L., Totzke, K., und Bickmore, T. (2015). Social support agents for older adults: longitudinal affective computing in the home. Journal on Multimodal User Interfaces, 9(1):79–88.

[Rist et al., 2015] Rist, T., Seiderer, A., Hammer, S., Mayr, M., und André, E. (2015). CARE - extending a digital picture frame with a recommender mode to enhance well-being of elderly people. In 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), Seiten 112–120.

[Rothman und Salovey, 1997] Rothman, A. J. und Salovey, P. (1997). Shaping perceptions to motivate healthy behavior: The role of message framing. Psychological Bulletin, 121:3–19.

[Rushton et al., 1987] Rushton, J., Murray, H. G., und Erdle, S. (1987). Combining trait consistency and learning specificity approaches to personality, with illustrative data on faculty teaching performance. Personality and Individual Differences, 8(1):59 – 66.

[Russell und Norvig, 2003] Russell, S. J. und Norvig, P. (2003). Artificial Intelligence: A modern approach. Prentice Hall, Upper Saddle River, NJ, 2nd int.. Auflage.

[Ryan und Deci, 2000] Ryan, R. M. und Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. Contemporary Educational Psychology, 25(1):54 – 67.

[Sarwar et al., 1998] Sarwar, B. M., Konstan, J. A., Borchers, A., Herlocker, J., Miller, B., und Riedl, J. (1998). Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. In Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work, CSCW '98, Seiten 345–354, New York, NY, USA. ACM.

[Satow, 2012] Satow, L. (2012). Big-Five-Persönlichkeitstest (B5T): Testmanual und Normen. `http://www.drsatow.de/tests/persoenlichkeitstest`.

[Schmid, 2005] Schmid, T., Herausgeber (2005). Promoting Health Through Creativity: For Professionals in Health, Arts and Education. Whurr Publishers.

[Schmidt und Birbaumer, 1996] Schmidt, R. F. und Birbaumer, N. (1996). Biologische Psychologie. Springer Verlag Berlin.

[Schwartz, 2004] Schwartz, S. H. (2004). Mapping and interpreting cultural differences around the world. In Vinken, H., Soeters, J., und Ester, P., Herausgeber, Comparing Cultures - Dimensions of Culture in Comparative Perspective, Seiten 43–73. Leiden, The Netherlands: Brill.

[Searle, 1969] Searle, J. R. (1969). Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press.

[Segmüller, 2015] Segmüller, F. (2015). Generierung überzeugender Empfehlungen und Argumente zur Förderung umweltbewussten Verhaltens basierend auf Kulturmodellen. Masterarbeit, betreut durch Hammer, S. und André, E., Universität Augsburg, Augsburg, Germany.

[Seiderer et al., 2015] Seiderer, A., Hammer, S., André, E., Mayr, M., und Rist, T. (2015). Exploring Digital Image Frames for Lifestyle Intervention to Improve Well-being of Older Adults. In Proc. of the 5th Int. Conf. on Digital Health 2015, DH '15, Seiten 71–78, New York, NY, USA. ACM.

[Sherchan et al., 2013] Sherchan, W., Nepal, S., und Paris, C. (2013). A survey of trust in social networks. ACM Comput. Surv., 45(4):47:1–47:33.

[Sherif et al., 1981] Sherif, C. W., Sherif, M., und Nebergall, R. E. (1981). Attitude and attitude change: The social judgment-involvement approach. Greenwood Press Westport, CT.

[Shigeyoshi et al., 2013] Shigeyoshi, H., Tamano, K., Saga, R., Tsuji, H., Inoue, S., und Ueno, T. (2013). Social Experiment on Advisory Recommender System for Energy-saving. In Proceedings of the 15th International Conference on Human Interface and the Management of Information: Information and Interaction Design - Volume Part I, HCI International'13, Seiten 545–554, Berlin, Heidelberg. Springer-Verlag.

[Simon et al., 2012] Simon, J., Jahn, M., und Al-Akkad, A. (2012). Saving energy at work: the design of a pervasive game for office spaces. In Proc. of the 11th Int. Conf. on Mobile and Ubiquitous Multimedia, MUM '12, Seiten 9:1–9:4, New York, NY, USA. ACM.

[Simons und Jones, 2011] Simons, H. W. und Jones, J. (2011). Persuasion in Society. Taylor & Francis.

[Sinha und Swearingen, 2002] Sinha, R. und Swearingen, K. (2002). The Role of Transparency in Recommender Systems. In CHI '02 Extended Abstracts on Human Factors in Computing Systems, CHI EA '02, Seiten 830–831, New York, NY, USA. ACM.

[Skevington et al., 2004] Skevington, S., Lotfy, M., und O'Connell, K. (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: Psychometric properties and results of the international field trial. A Report from the WHOQOL Group. Quality of Life Research, 13(2):299–310.

[Srinivasan und Takayama, 2016] Srinivasan, V. und Takayama, L. (2016). Help Me Please: Robot Politeness Strategies for Soliciting Help From Humans. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, Seiten 4945–4955, New York, NY, USA. ACM.

[Steghöfer et al., 2010] Steghöfer, J., Kiefhaber, R., Leichtenstern, K., Bernard, Y., Klejnowski, L., Reif, W., Ungerer, T., André, E., Hähner, J., und Müller-Schloer, C. (2010). Trustworthy Organic Computing Systems: Challenges and Perspectives. In Xie, B., Branke, J., Sadjadi, S. M., Zhang, D., und Zhou, X., Herausgeber, Autonomic and Trusted Computing - 7th International Conference, ATC 2010, Xi'an, China, October 26-29, 2010. Proceedings, Band 6407 in Lecture Notes in Computer Science, Seiten 62–76. Springer.

[Steinmetz, 1999] Steinmetz, G. (1999). State/Culture: State-formation after the cultural turn. Cornell University Press.

[Stephenson et al., 2010] Stephenson, J., Barton, B., Carrington, G., Gnoth, D., Lawson, R., und Thorsnes, P. (2010). Energy cultures: A framework for understanding energy behaviours. Energy Policy, 38(10):6120 – 6129. The socio-economic transition towards a hydrogen economy - findings from European research, with regular papers.

[Strait et al., 2014] Strait, M., Canning, C., und Scheutz, M. (2014). Let Me Tell You! Investigating the Effects of Robot Communication Strategies in Advice-giving Situations Based on Robot Appearance, Interaction Modality and Distance. In Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction, HRI '14, Seiten 479–486, New York, NY, USA. ACM.

[Strom-Magazin.de, 2017] Strom-Magazin.de (2017). Der große Heizungsvergleich: Kosten, Nutzen, Effizienz. `https://www.strom-magazin.de/heizung-vergleich`.

[Svennevig, 1999] Svennevig, J. (1999). Getting Acquainted in Conversation: A Study of Initial Interactions (Pragmatics & beyond, new ser. 64). John Benjamins Publishing Company.

[Swearingen und Sinha, 2002] Swearingen, K. und Sinha, R. (2002). Interaction Design for Recommender Systems. In In Designing Interactive Systems 2002. ACM. Press.

[Teri und Lewinsohn, 1982] Teri, L. und Lewinsohn, P. (1982). Modification of the Pleasant and Unpleasant Events Schedules for use with the elderly. Journal of consulting and clinical psychology, 50(3):444–445.

[Tintarev und Masthoff, 2008] Tintarev, N. und Masthoff, J. (2008). The Effectiveness of Personalized Movie Explanations: An Experiment Using Commercial Meta-data. In Nejdl, W., Kay, J., Pu, P., und Herder, E., Herausgeber, Adaptive Hypermedia and Adaptive Web-Based Systems, Band 5149 in Lecture Notes in Computer Science, Seiten 204–213. Springer Berlin Heidelberg.

[Tintarev und Masthoff, 2011] Tintarev, N. und Masthoff, J. (2011). Designing and Evaluating Explanations for Recommender Systems. In Ricci, F., Rokach, L., Shapira, B., und Kantor, B. P., Herausgeber, Recommender Systems Handbook, Seiten 479–510. Springer US, Boston, MA.

[Torrey et al., 2013] Torrey, C., Fussell, S., und Kiesler, S. (2013). How a Robot Should Give Advice. In Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction, HRI '13, Seiten 275–282, Piscataway, NJ, USA. IEEE Press.

[Triandis, 1995] Triandis, H. C. (1995). Individualism & Collectivism. Westview Press.

[Turunen et al., 2008] Turunen, M., Hakulinen, J., Smith, C., Charlton, D., und Zhang, L. (2008). Physically Embodied Conversational Agents as Health and Fitness Companions. In Proceedings of the 9th Annual Conference of the International Speech Communication Association, INTERSPEECH 2008, Seiten 2466–2469. ISCA.

[van Mulken et al., 1999] van Mulken, S., André, E., und Müller, J. (1999). An Empirical Study on the Trustworthiness of Life-like Interface Agents. In Proceedings of the 8th International Conference on Human-Computer Interaction (HCI International 1999), Seiten 152–156, Hillsdale, NJ, USA. L. Erlbaum Associates Inc.

[van Pinxteren et al., 2011] van Pinxteren, Y., Geleijnse, G., und Kamsteeg, P. (2011). Deriving a Recipe Similarity Measure for Recommending Healthful Meals. In Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI '11, Seiten 105–114, New York, NY, USA. ACM.

[Venkatesh und Davis, 2000] Venkatesh, V. und Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Manage. Sci., 46(2):186–204.

[Venkatesh et al., 2003] Venkatesh, V., Morris, M. G., Davis, G. B., und Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. MIS Q., 27(3):425–478.

[Wagner et al., 2010] Wagner, N., Hassanein, K., und Head, M. (2010). Review: Computer Use by Older Adults: A Multi-disciplinary Review. Comput. Hum. Behav., 26(5):870–882.

[Wang und Benbasat, 2008] Wang, W. und Benbasat, I. (2008). Attributions of Trust in Decision Support Technologies: A Study of Recommendation Agents for E-Commerce. J. Manage. Inf. Syst., 24(4):249–273.

[Williams et al., 2006] Williams, J. H., Auslander, W. F., de Groot, M., Robinson, A. D., Houston, C., und Haire-Joshu, D. (2006). Cultural Relevancy of a Diabetes Prevention Nutrition Program for African American Women. Health Promotion Practice, 7(1):56–67.

[Wilson et al., 2009] Wilson, D. C., Leland, S., Godwin, K., Baxter, A., Levy, A., Smart, J., Najjar, N., und Andaparambil, J. (2009). SmartChoice: An Online Recommender System to Support Low-Income Families in Public School Choice. AI Magazine, 30(2):46–58.

[Wissner et al., 2014] Wissner, M., Hammer, S., Kurdyukova, E., und André, E. (2014). Trust-based Decision-making for the Adaptation of Public Displays in Changing Social Contexts. Journal of Trust Management, 1(1):6.

[Wood, 2000] Wood, W. (2000). Attitude Change: Persuasion and Social Influence. Annual Review of Psychology, 51(1):539–570.

[Wu et al., 2011] Wu, K., Zhao, Y., Zhu, Q., Tan, X., und Zheng, H. (2011). A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. International Journal of Information Management, 31(6):572 – 581.

[WWF Deutschland, 2016] WWF Deutschland (2016). Energie sparen - praktische Tipps für Ihren Haushalt. `http://www.wwf.de/aktiv-werden/tipps-fuer-den-alltag/energie-spartipps/strom-sparen/`.

[Yan und Holtmanns, 2008] Yan, Z. und Holtmanns, S. (2008). Trust Modeling and Management: From Social Trust to Digital Trust, Kapitel 18, Seiten 279–303. IGI Global.

[Yu et al., 2017] Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., und Chen, F. (2017). User Trust Dynamics: An Investigation Driven by Differences in System Performance. In Proceedings of the 22Nd International Conference on Intelligent User Interfaces, IUI '17, Seiten 307–317, New York, NY, USA. ACM.

[Zanker, 2012] Zanker, M. (2012). The Influence of Knowledgeable Explanations on Users' Perception of a Recommender System. In Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys '12, Seiten 269–272, New York, NY, USA. ACM.

A Theoriebasierte Empfehlungsauswahl - Ergebnistabellen

A.1 Vergleich der Empfehlungsqualitäten im Sparsity-Szenario (CARE)

SG	MAE				F1-Maß			
	BB	WB	ME	LIN	BB	WB	ME	LIN
25	0,590	0,629 (-6,6%)	0,601 (-1,7%)	0,594 (-0,6%)	0,522	**0,468 (-10,4%)**	0,513 (-1,7%)	0,520 (-0,3%)
30	0,579	0,622 (-7,4%)	0,602 (-3,9%)	0,585 (-1,0%)	0,530	**0,477 (-10,0%)**	0,523 (-1,4%)	0,525 (-0,9%)
35	0,599	0,630 (-5,2%)	0,608 (-1,7%)	0,599 (-0,1%)	0,519	0,476 (-8,3%)	0,510 (-1,6%)	0,519 (0,0%)
40	0,602	0,620 (-2,9%)	0,610 (-1,4%)	0,603 (-0,1%)	0,519	0,480 (-7,6%)	0,514 (-0,9%)	0,515 (-0,9%)
45	0,615	0,630 (-2,4%)	0,623 (-1,3%)	0,613 (0,4%)	0,512	0,486 (-5,0%)	0,515 (0,6%)	0,510 (-0,3%)
50	0,614	0,631 (-2,8%)	0,621 (-1,2%)	0,617 (-0,4%)	0,507	0,479 (-5,6%)	0,504 (-0,6%)	0,503 (-0,7%)
55	0,628	0,630 (-0,4%)	0,625 (0,4%)	0,617 (1,7%)	0,505	0,483 (-4,3%)	0,505 (-0,1%)	0,503 (-0,4%)
60	0,656	0,641 (2,3%)	0,641 (2,3%)	0,634 (3,4%)	0,500	0,479 (-4,0%)	0,497 (-0,5%)	0,501 (0,3%)
65	0,682	0,659 (3,3%)	0,657 (3,6%)	0,652 (4,3%)	0,486	0,472 (-2,7%)	0,485 (-0,2%)	0,483 (-0,5%)
70	0,714	0,661 (7,5%)	0,661 (7,4%)	0,657 (8,0%)	0,479	0,472 (-1,4%)	0,488 (1,9%)	0,481 (0,5%)
75	0,783	**0,694 (11,4%)**	**0,700 (10,7%)**	**0,702 (10,4%)**	0,452	0,450 (-0,5%)	0,456 (0,9%)	0,456 (0,7%)
80	0,832	**0,721 (13,4%)**	**0,728 (12,4%)**	**1,043 (12,6%)**	0,451	0,448 (-0,7%)	0,521 (1,5%)	0,456 (1,0%)
85	0,894	**0,769 (14,1%)**	**0,776 (12,6%)**	**1,108 (13,2%)**	0,427	0,439 (2,8%)	0,497 (2,8%)	0,439 (2,7%)
90	0,915	**0,810 (11,5%)**	**0,824 (10,0%)**	**0,814 (11,0%)**	0,404	0,431 (6,5%)	0,471 (7,1%)	0,429 (6,2%)
95	0,895	0,868 (3,0%)	0,867 (3,1%)	0,867 (3,1%)	0,394	0,365 (5,1%)	0,384 (5,2%)	0,380 (4,2%)

SG = Grad der Spärlichkeit; BB = Bewertungsbasierter kollaborativer Filter; WB = Theoriebasierter Filter mit Wohlbefindens-Modell; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter

A.2 Vergleich der Empfehlungsqualitäten im New-User-Szenario (CARE)

VB	MAE				F1-Maß			
	BB	WB	ME	LIN	BB	WB	ME	LIN
1	0,984	**0,883 (10,3%)**	**0,884 (10,2%)**	**0,885 (10,1%)**	0,393	**0,435 (10,5%)**	**0,437 (11,3%)**	**0,432 (10,0%)**
2	0,860	**0,743 (13,6%)**	**0,746 (13,2%)**	**0,747 (13,1%)**	0,436	0,457 (4,7%)	0,450 (3,2%)	0,459 (5,3%)
3	0,780	0,705 (9,6%)	0,707 (9,4%)	0,709 (9,2%)	0,457	0,461 (0,9%)	0,459 (0,4%)	0,462 (1,1%)
4	0,730	0,693 (5,1%)	0,692 (5,2%)	0,694 (4,9%)	0,470	0,460 (-2,1%)	0,457 (-2,7%)	0,465 (-1,1%)
5	0,708	0,682 (3,7%)	0,684 (3,4%)	0,683 (3,6%)	0,467	0,454 (-2,8%)	0,454 (-2,6%)	0,460 (-1,4%)

VB = Anzahl vorliegender Bewertungen; BB = Bewertungsbasierter kollaborativer Filter; WB = Theoriebasierter Filter mit Wohlbefindens-Modell; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter

A.3 Vergleich der Empfehlungsqualitäten im Sparsity-Szenario (SavER)

SG	MAE				F1-Maß			
	BB	EC	ME	LIN	BB	EC	ME	LIN
25	0,853	0,895 (-4,8%)	0,858 (-0,6%)	0,851 (0,3%)	0,546	0,533 (-2,4%)	0,563 (3,1%)	0,548 (0,4%)
30	0,864	0,901 (-4,3%)	0,878 (-1,7%)	0,865 (-0,1%)	0,562	0,551 (-1,9%)	0,586 (4,3%)	0,563 (0,3%)
35	0,862	0,896 (-4,0%)	0,877 (-1,7%)	0,866 (-0,5%)	0,574	0,552 (-3,8%)	0,582 (1,5%)	0,577 (0,7%)
40	0,883	0,908 (-2,8%)	0,891 (-0,9%)	0,881 (0,3%)	0,567	0,553 (-2,5%)	0,584 (3,0%)	0,569 (0,3%)
45	0,901	0,918 (-2,0%)	0,904 (-0,4%)	0,900 (0,1%)	0,567	0,552 (-2,6%)	0,584 (3,1%)	0,567 (0,1%)
50	0,897	0,915 (-2,1%)	0,903 (-0,7%)	0,895 (0,2%)	0,579	0,566 (-2,2%)	0,587 (1,3%)	0,579 (0,0%)
55	0,928	0,927 (0,1%)	0,920 (0,9%)	0,918 (1,1%)	0,557	0,552 (-1,0%)	0,571 (2,4%)	0,561 (0,7%)
60	0,958	0,942 (1,6%)	0,940 (1,9%)	0,937 (2,2%)	0,548	0,545 (-0,6%)	0,567 (3,5%)	0,556 (1,4%)
65	1,008	0,972 (3,6%)	0,970 (3,7%)	0,966 (4,2%)	0,541	0,540 (-0,1%)	0,559 (3,3%)	0,545 (0,7%)
70	1,054	0,988 (6,2%)	0,991 (5,9%)	0,985 (6,5%)	0,519	0,527 (1,6%)	0,546 (5,3%)	0,526 (1,4%)
75	1,102	1,006 (8,7%)	1,008 (8,5%)	1,007 (8,6%)	0,513	0,522 (1,7%)	0,539 (5,0%)	0,523 (1,9%)
80	1,182	**1,039 (12,0%)**	**1,042 (11,8%)**	**1,043 (11,7%)**	0,498	0,514 (3,2%)	0,521 (4,8%)	0,510 (2,4%)
85	1,259	**1,104 (12,3%)**	**1,118 (11,2%)**	**1,108 (12,0%)**	0,469	0,488 (4,0%)	0,497 (6,0%)	0,482 (2,7%)
90	1,298	1,195 (8,0%)	1,201 (7,5%)	1,198 (7,7%)	0,447	0,463 (3,6%)	0,471 (5,5%)	0,460 (3,0%)
95	1,281	1,275 (0,4%)	1,274 (0,5%)	1,275 (0,5%)	0,394	0,399 (1,3%)	0,399 (1,1%)	0,399 (1,1%)

SG = Grad der Spärlichkeit; BB = Bewertungsbasierter kollaborativer Filter; EC = Theoriebasierter Filter mit Energiekulturen; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter

A.4 Vergleich der Empfehlungsqualitäten im New-User-Szenario (SavER)

VB	MAE				F1-Maß			
	BB	EC	ME	LIN	BB	EC	ME	LIN
1	1,372	**1,201 (12,5%)**	**1,200 (12,5%)**	**1,204 (12,3%)**	0,419	**0,476 (13,6%)**	**0,474 (13,2%)**	**0,474 (13,0%)**
2	1,156	1,053 (8,9%)	1,054 (8,8%)	1,055 (8,7%)	0,490	0,501 (2,2%)	0,504 (2,8%)	0,511 (4,3%)
3	1,027	0,992 (3,4%)	0,993 (3,3%)	0,991 (3,5%)	0,527	0,522 (-1,1%)	0,526 (-0,3%)	0,536 (1,7%)
4	0,985	0,974 (1,1%)	0,970 (1,5%)	0,970 (1,5%)	0,537	0,534 (-0,6%)	0,537 (0,0%)	0,543 (1,0%)
5	0,957	0,956 (0,2%)	0,950 (0,8%)	0,948 (1,0%)	0,549	0,540 (-1,5%)	0,550 (0,3%)	0,556 (1,4%)

VB = Anzahl vorliegender Bewertungen; BB = Bewertungsbasierter kollaborativer Filter; EC = Theoriebasierter Filter mit Energiekulturen; ME = Hybrider Filter mit Merkmalserweiterung; LIN = Linearer hybrider Filter

B Persönlichkeitstest nach Satow

Alle 50 Aussagen werden auf einer Skala von 1 = „trifft gar nicht zu“ bis 4 = „trifft genau zu“ bewertet. Bei negativ gepolten Aussagen (-) verläuft die Punktevergabe bei der Auswertung von 4 bis 1. Eine ausführliche Dokumentation des Tests ist online zu finden [Satow, 2012].

Neurotizismus

1. Ich bin ein ängstlicher Typ.
2. Ich fühle mich oft unsicher.
3. Ich verspüre oft eine große innere Unruhe.
4. Ich mache mir oft unnütze Sorgen.
5. Ich grübele viel über meine Zukunft nach.
6. Oft überwältigen mich meine Gefühle.
7. Ich bin oft ohne Grund traurig.
8. Ich bin oft nervös.
9. Oft werde ich von meinen Gefühlen hin- und hergerissen.
10. Ich bin mir in meinen Entscheidungen oft unsicher

Extraversion

1. Ich bin gerne mit anderen Menschen zusammen.
2. Ich kann schnell gute Stimmung verbreiten.
3. Ich bin unternehmungslustig.
4. Ich stehe gerne im Mittelpunkt.
5. Im Grunde bin ich oft lieber für mich allein. (-)
6. Ich bin ein Einzelgänger. (-)
7. Ich gehe gerne auf Partys.
8. Ich bin in vielen Vereinen aktiv.
9. Ich bin ein gesprächiger und kommunikativer Mensch.
10. Ich bin sehr kontaktfreudig.

Gewissenhaftigkeit

1. Ich bin sehr pflichtbewusst.
2. Meine Aufgaben erledige ich immer sehr genau.
3. Ich war schon als Kind sehr ordentlich.

4. Ich gehe immer planvoll vor.
5. Ich habe meine festen Prinzipien und halte daran auch fest.
6. Auch kleine Bußgelder sind mir sehr unangenehm.
7. Auch kleine Schlampereien stören mich.
8. Ich achte sehr darauf, dass Regeln eingehalten werden.
9. Wenn ich mich einmal entschieden habe, dann weiche ich davon auch nicht mehr ab.
10. Ich mache eigentlich nie Flüchtigkeitsfehler.

Offenheit

1. Ich will immer neue Dinge ausprobieren.
2. Ich bin ein neugieriger Mensch.
3. Ich reise viel, um andere Kulturen kennenzulernen.
4. Am liebsten ist es mir, wenn alles so bleibt, wie es ist. (-)
5. Ich diskutiere gerne.
6. Ich lerne immer wieder gerne neue Dinge.
7. Ich beschäftige mich viel mit Kunst, Musik und Literatur.
8. Ich interessiere mich sehr für philosophische Fragen.
9. Ich lese viel über wissenschaftliche Themen, neue Entdeckungen oder historische Begebenheiten.
10. Ich habe viele Ideen und viel Fantasie.

Verträglichkeit

1. Ich achte darauf, immer freundlich zu sein.
2. Ich bin ein höflicher Mensch.
3. Ich helfe anderen, auch wenn man mir es nicht dankt.
4. Ich habe immer wieder Streit mit anderen. (-)
5. Ich bin ein Egoist. (-)
6. Wenn mir jemand hilft, erweise ich mich immer als dankbar.
7. Ich würde meine schlechte Laune nie an anderen auslassen.
8. Es fällt mir sehr leicht, meine Bedürfnisse für andere zurückzustellen.
9. Ich kann mich gut in andere Menschen hinein versetzen.
10. Ich komme immer gut mit anderen aus, auch wenn sie nicht meiner Meinung sind.

C Liste eigener Publikationen

Referenzen	Art der Publikation	Eigener Beitrag	Thema bzw. Beitrag zur Dissertation
[Hammer et al., 2010]	Konferenz	Erstautor; Implementierung (Rule Engine)	Empfehlungssystem für Diabetiker (Kapitel 1.1)
[Kiefhaber et al., 2011]	Workshop	Co-Autor; Konzept (Reputationsmetrik)	Evaluation von Reputationsmetriken
[Kurdyukova et al., 2012]	Konferenz	Co-Autor; Konzept und Implementierung (Empfehlungssystem); Nutzerstudie	UX proaktiver Empfehlungssysteme
[Bühling et al., 2012]	Konferenz	Co-Autor; Implementierung (mobile App); Nutzerstudie	Mobiles AR-Quiz zum Thema Energiesparen
[Bee et al., 2012]	Konferenz	Co-Autor; Konzept (Nutzerstudie)	Entwicklung des UTM (Kapitel 6.2)
[Hammer et al., 2013]	Workshop	Erstautor; Nutzerstudie	UX von Reputationssystemen
[Kurdyukova et al., 2013]	Konferenz	Co-Autor; Konzept (Nutzerstudie)	Anwendungsszenario für UTM (Öffentliche Displays)
[Anders et al., 2013]	Technischer Report	Co-Autor; Beschreibung Referenzarchitektur	Referenzarchitektur eines Multi-Display-Multi-Nutzer Systems das Nutzervertrauen berücksichtigt
[Hammer et al., 2014]	Konferenz	Erstautor; Implementierung (SavER-System); Nutzerstudie	Anwendungsszenario für UTM (SavER Kapitel 6)
[Wissner et al., 2014]	Journal	Co-Autor; Konzept (Nutzerstudie)	Anwendungsszenario für UTM (Öffentliche Displays)
[Hammer et al., 2015a]	Workshop	Erstautor; Nutzerstudie	Effekt der Energiekultur auf Präferenzen für Energiespartipps (Kapitel 4.3.3)
[Hammer et al., 2015b]	Konferenz	Erstautor; Nutzerstudie	Anforderungsanalyse für CARE (Kapitel 4.4)

Referenzen	Art der Publikation	Eigener Beitrag	Thema bzw. Beitrag zur Dissertation
[Hammer et al., 2015c]	Journal	Erstautor; Implementierung (SavER-System); Nutzerstudie	Anwendungsszenario für UTM (SavER - Kapitel 6)
[Rist et al., 2015]	Konferenz	Co-Autor; Implementierung (CARE-Empfehlungsauswahl); Nutzerstudie	CARE-Prototyp und Evaluation (Kapitel 4.4)
[Seiderer et al., 2015]	Konferenz	Co-Autor; Implementierung (CARE-Empfehlungsauswahl); Nutzerstudie	CARE-Prototyp und Evaluation (Kapitel 4.4)
[Bogomolov, 2015]	Masterarbeit	Betreuer	Evaluation von Höflichkeitsstrategien (Kapitel 5.2)
[Segmüller, 2015]	Masterarbeit	Betreuer	Effekt der Energiekultur auf Präferenzen für Energiespartipps (Kapitel 4.3.3); Kulturabhängige Argumente (Kapitel 5.1)
[Mertens, 2016]	Masterarbeit	Betreuer	Persönlichkeitsausprägungen von Empfehlungstexten (Kapitel 5.3)
[Hammer et al., 2016a]	Konferenz	Erstautor; Nutzerstudie	Evaluation von Höflichkeitsstrategien (Kapitel 5.2)
[Hammer et al., 2016b]	Buchkapitel	Erstautor; Implementierung (SavER-System); Nutzerstudie	Zusammenfassung OCTrust Projekt (u.a. SavER Kapitel 6)
[Hammer et al., 2017] (Nominierung Best Late Breaking Report)	Konferenz	Erstautor; Nutzerstudie	Vergleich von sozialen Robotern und Tablet PCs als Interaktionsgerät (CARE)